

Data Analytics with Python

Prof. A. Ramesh
Computer Science and
Engineering
IIT Roorkee



INDEX

S. No	Topic	Page No
	<i>Week 1</i>	
1	Introduction to data analytics	1
2	Python Fundamentals - I	33
3	Python Fundamentals - II	54
4	Central Tendency and Dispersion - I	83
5	Central Tendency and Dispersion - II	108
	<i>Week 2</i>	
6	Introduction to Probability- I	127
7	Introduction to Probability- II	155
8	Probability Distributions - I	177
9	Probability Distributions - II	198
10	Probability Distributions - III	225
	<i>Week 3</i>	
11	Python Demo for Distributions	246
12	Sampling and Sampling Distribution	256
13	Distribution of Sample Means, population, and variance	287
14	Confidence interval estimation: Single population - I	304
15	Confidence interval estimation: Single population - II	324
	<i>Week 4</i>	
16	Hypothesis Testing- I	342
17	Hypothesis Testing- II	364
18	Hypothesis Testing- III	380
19	Errors in Hypothesis Testing	394
20	Hypothesis Testing: Two sample test- I	422
	<i>Week 5</i>	
21	Hypothesis Testing: Two sample test- II	442
22	Hypothesis Testing: Two sample test- III	464
23	ANOVA - I	480
24	ANOVA - II	494
25	Post Hoc Analysis(Tukeyâ€™s test)	513
	<i>Week 6</i>	

26	Randomize block design (RBD)	542
27	Two Way ANOVA	563
28	Linear Regression - I	583
29	Linear Regression - II	601
30	Linear Regression - III	614

Week 7

31	Estimation, Prediction of Regression Model Residual Analysis	634
32	Estimation, Prediction of Regression Model Residual Analysis - II	652
33	Multiple Regression Model - I	674
34	Multiple Regression Model-II	695
35	Categorical variable regression	714

Week 8

36	Maximum Likelihood Estimation- I	744
37	Maximum Likelihood Estimation-II	761
38	Logistic Regression- I	785
39	Logistic Regression-II	802
40	Linear Regression Model Vs Logistic Regression Model	818

Week 9

41	Confusion matrix and ROC- I	838
42	Confusion Matrix and ROC-II	860
43	Performance of Logistic Model-III	883
44	Regression Analysis Model Building - I	895
45	Regression Analysis Model Building (Interaction)- II	910

Week 10

46	Chi - Square Test of Independence - I	928
47	Chi-Square Test of Independence - II	949
48	Chi-Square Goodness of Fit Test	971
49	Cluster analysis: Introduction- I	990
50	Clustering analysis: part II	1009

Week 11

51	Clustering analysis: Part III	1026
52	Cluster analysis: Part IV	1046
53	Cluster analysis: Part V	1068

54	K- Means Clustering	1083
55	Hierarchical method of clustering -I	1109

Week 12

56	Hierarchical method of clustering- II	1134
57	Classification and Regression Trees (CART : I)	1162
58	Measures of attribute selection	1187
59	Attribute selection Measures in CART : II	1206
60	Classification and Regression Trees (CART) - III	1224

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology, Roorkee

Lecture No 1
Introduction to Data Analytics

Welcome students this course on data analytics with the Python today is the introduction class. This lecture is on introduction to data analytics.

(Refer Slide Time: 00:34)

Objective of the course

- The principle focus of this course is to introduce conceptual understanding using simple and practical examples rather than repetitive and point click mentality
- This course should make you comfortable using analytics in your career and your life
- You will know how to work with real data, and might have learned many different methodologies but choosing the right methodology is important

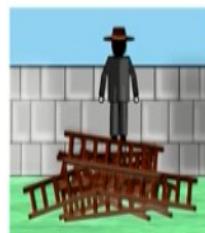
The objective of this course is to introduce the conceptual understanding using simple and practical examples rather than repetitive and point clique mentality, here most of the students generally they are, how they are using the software for doing data analytics. Just they want to just click it, they want to get the result, they do not want bother about exactly what is happening inside the software. This course should make you comfortable using analytics in your career and your life.

You will know how to work with a real data and you might have learnt the many different methodologies, but choosing the right methodology is important. This course will focus you will help you how to choose the right data analytical tools.

(Refer Slide Time: 01:17)

Objective of the course Contd...

- The danger in using quantitative method does not generally lie in the inability to perform the calculation
- The real threat is lack of fundamental understanding of:
 - Why to use a particular technique or procedure
 - How to use it correctly and,
 - How to correctly interpret the result



Objective of the course is, when you look at this picture. How this person is using this tool, there is a ladder. He was not knowing correctly how to use this ladder for the purpose it is intended. So the danger in using quantitative method does not generally lie in the inability to perform the calculation, because of the computer development in computer technology.

There are many packages available for doing data analytics. But, the real threat is lack of fundamental understanding of why to use particular technique or procedures and how to use it correctly and, how to correctly interpret the result. This course will focus on how to choose the right technique and how to use it correctly and how to interpret the result.

(Refer Slide Time: 02:01)

Learning objectives

1. Define data and its importance
2. Define data analytics and its types
3. Explain why analytics is important in today's business environment
4. Explain how statistics, analytics and data science are interrelated
5. Why python?
6. Explain the four different levels of Data:
 - Nominal
 - Ordinal
 - Interval and
 - Ratio

So what was the learning objective of this class that is; after completing this lecture what you will learn one is you can define what is data and its importance. You can define what are data analytics and types. You can explain why analytics is in today's business environment is so important. Then we can see how statistics, analytics and data science are interrelated, there seems to be some overlap in this we will clarify that what is the difference, how these are overlapped how these are interrelated.

In this course we are going to use a package called Python. I will explain how and why it is important to use the Python in this course, at the end of this session we will explain the four important levels of data that is nominal, ordinal, interval and ratio. Now we will go to the content;

(Refer Slide Time: 02:54)

1. Define Data and its importance

We will define data and its importance. There are three term Variable, Measurement and Data. One is variable, What is generating so much data? measurement and data. Next we will see what is generating so much data. Next we will see why data is important? how data add value to the business, and then we will say why data is important.

o

(Refer Slide Time: 03:11)

1.1 Variable, Measurement and Data

- Variables – is a characteristic of any entity being studied that is capable of taking on different values
- Measurements – is when a standard process is used to assign numbers to particular attributes or characteristic of a variable
- Data – data are recorded measurements

0

See the variable, measurement and data these are the terms which we are going to use frequently in this course. So what is a variable? Variable is a characteristic of any entity being studied that is capable of taking on different values. Say for example, X is the variable it can take any values it may be 1 it may be 2 or it may be 0 and so on. The measurement is, when you standard processes used to assign numbers to your particular attributes or characteristics of variable are called a measurement.

For that X, you want to substitute some values. For that value, you have to measure the characteristics of the variable, that is nothing but your measurement. So then, what is the data? Data are recorded measurement. So there is a variable you measure the phenomena, after measuring the phenomena you are substituting some value for the variable so the variable will take a particular value that value is nothing but your data.

So X is the variable for example number 5 is the data. How you are measuring that 5, that is called measurement. Then what is generating so much of data.

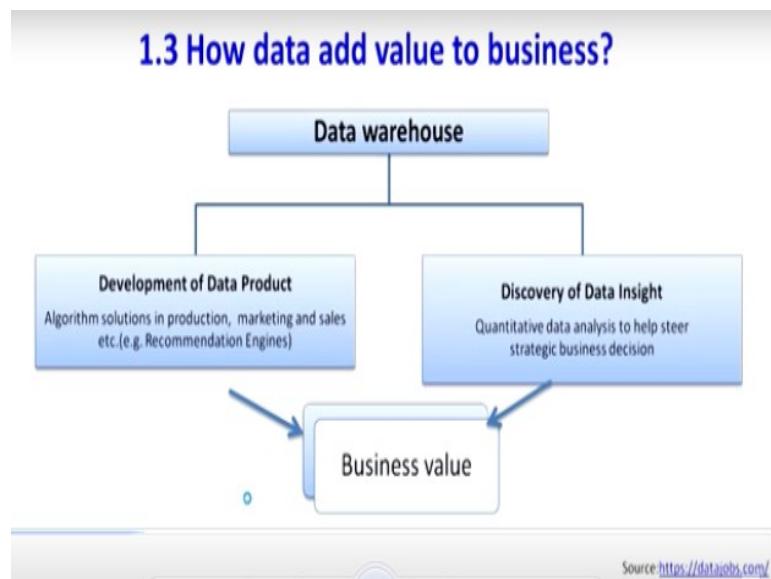
(Refer Slide Time: 04:33)

1.2 What is generating so much data?

- Data can be generated by
 - Humans,
 - Machines or
 - Humans-machines combines
- It can be generated anywhere where any information is generated and stored in structured or unstructured formats

Data can be generated different way humans, machines, and human - machines combines. The humans, machines and human - machines combines in the sense, now seen everybody is having the various Facebook account, we have LinkedIn account, we are in various social network sites. Now the availability of the data is not the problem. It can be generated anywhere where the information is generated and stored and structured or unstructured format.

(Refer Slide Time: 05:06)



So how the data add value to the business? So the data after getting from various sources assume that it is a store in the form of data warehouse. So from the data warehouse the data can be used for development of a data product. Here we are using the word data product and in the coming slides, I will explain exactly what is the data product with some examples. So

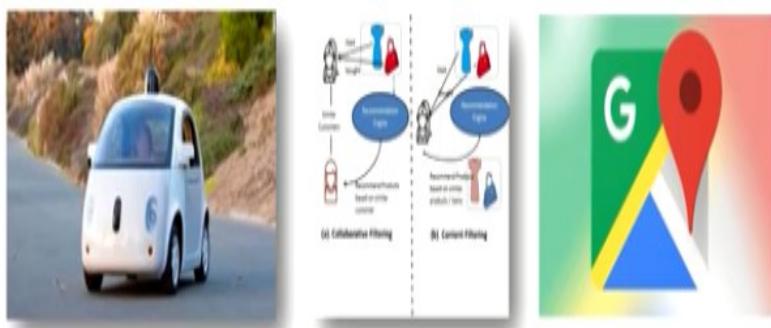
the same data, if we look at the right hand side that can be used to get more insights from the data.

Okay, what do you mean the data product? For example, algorithm solutions in production, marketing and sales, example of some data product. For example, recommendation engine one of the example for data product. Suppose, if you go for Flipkart or Amazon for buying a particular product in that package, that software itself, we will recommend to you what is the next product, possible product that you can buy. That is nothing but the recommendation engine.

Even if you watch some YouTube videos on particular topic, that YouTube itself will suggest to you what are the relevant videos are available. So that is a recommendation engine. That is one of the examples of your data product; with the help of data so that will help you to forming a data product or you can get an insight from the data. That will add your business value to you.

(Refer Slide Time: 06:27)

Data Products



See this is an example of your data products, this is the driverless car. Google car, so the whole concept of Google car is with the help of data. It is detecting all other requirements for driving the cars. The next example is for recommendation engine, as I told you previously when you buy any product they will suggest you that along with this product, the other product also can be purchased.

Another very common example for a data product is Google. The Google has lot of applications, one of the application of example for data product is Google mapping. So the Google mapping is helping you to find out what is the right route, which road there is a traffic, in which road there is a toll booth, so this kind of information we can get it from the Google map. So this Google map is the one of an example of your data production.

(Refer Slide Time: 07:20)

1.4 Why Data is important?

- Data helps in make better decisions
- Data helps in solve problems by finding the reason for underperformance
- Data helps one to evaluate the performance.
- Data helps one improve processes
- Data helps one understand consumers and the market

Now why data is important? The data helps in making better decisions, data helps in solve problem by finding the reason for underperformance. Suppose some company it is not performing properly by collecting the data we can identify what was the reason for this under performance. The data helps one to evaluate the performance. So what is the current performance, the data also can be used for benchmarking the performance of your business organization.

And after benchmarking data helps one improving the performance also, so data also can help one understand the consumers and the markets, especially the marketing context. You can understand who are the right consumers and what kind of preferences they are having in the market.

(Refer Slide Time: 08:16)

2. Define data analytic and its types

- Define data analytics
- Why analytics is important?
- Data analysis
- Data analytics vs. Data analysis
- Types of Data analytics

Next we will define what is a data analytics and its types? So in this coming two, three slides we are going to discuss, we will define what is data analytics? Then you say why analytics is important? Then we will see that data analysis? Then we will see how data analytics is different from data analysis? At the end will we see types of data analytics?

(Refer Slide Time: 08:40)

2.1. Define data analytics

- Analytics is defined as “the scientific process of transforming data into insights for making better decisions”
- Analytics, is the use of data, information technology, statistical analysis, quantitative methods, and mathematical or computer-based models to help managers gain improved insight about their business operations and make better, fact-based decisions – James Evans
- Analysis = Analytics ?

We will define data analytics is the scientific process of transforming data into insights for making better decisions. See it is a scientific process for transforming the data into for making better decisions, even without the data also even without doing analytics also you can make the decision but you cannot make the better decision without analytics. By the virtue of your experience on intuitions you can take the decisions that also sometimes may be correct.

But about the help of data if you are making the decision then that will enable you to make the better decisions. Another professor James Evans, he has defined the data analytics in this way. it is the use of the data information technology, statistical analysis, quantitative methods and mathematical or computer-based models to help managers gain improved insight about their business operations and make better, fact-based decisions.

You see that there are many terms which are appearing here, one is IT, next one is a statistical analysis, and next one is the quantitative methods, then mathematical knowledge and computer-based models. So when we will see how these are interrelated in coming slides. Generally, among the students, there is a confusion whether the analysis and analytics is same or different?

(Refer Slide Time: 10:13)

2.2 Why analytics is important?

- Opportunity abounds for the use of analytics and big data such as:
 1. Determining credit risk
 2. Developing new medicines
 3. Finding more efficient ways to deliver products and services
 4. Preventing fraud
 5. Uncovering cyber threats
 6. Retaining the most valuable customers



Why analytics is important. The opportunity abounds for the use of analytics and big data such as: for determining the credit risk, for developing new medicines, especially in healthcare. The healthcare analytics is an emerging, that is helping you to identify what is the correct medicines. Finding more efficient ways to deliver product and services. For example: in the banking context data analytics is used for preventing the fraud, and it is uncovering the cyber threats.

With the help of data analytics you can find out the possible cyber crimes and we can detect it we can prevent it. And data analytics are also important for retaining the most valuable customers. We can identify who is your valuable customer or non valuable customers. So we can focus on more on our valuable customers. Okay,

(Refer Slide Time: 11:08)

2.3 Data analysis

- Data analysis is the process of examining, transforming, and arranging raw data in a specific way to generate useful information from it
- Data analysis allows for the evaluation of data through analytical and logical reasoning to lead to some sort of outcome or conclusion in some context
- Data analysis is a multi-faceted process that involves a number of steps, approaches, and diverse techniques

Now what is the data analysis? Is the process of examining, transforming and arranging raw data in a specific way to generate useful information from it. So data analysis allows for the evaluation of data through analytical and logical reasoning to lead to some sort of outcome or conclusion in some context. Data analysis is a multi-faceted process that involves a number of steps approaches and diverse techniques. That we will see in coming lecture.

(Refer Slide Time: 11:41)

2.4 Data analytics vs. Data analysis



So now we will see what is the analysis is data analysis and data analytics. When you say analysis when you say data analysis it is something about what has happened in the past. So we will explain why that has happened? We will explain how it has happened? We can explain why it has happened? For example, when we say data analysis that is nothing about

studying about what has happened it is like kind of a post-mortem analysis. What has happened in the past?

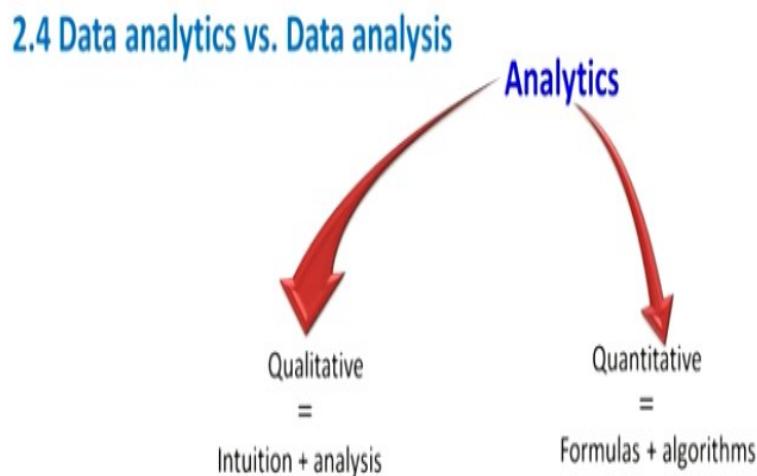
(Refer Slide Time: 12:13)



Explore potential future events

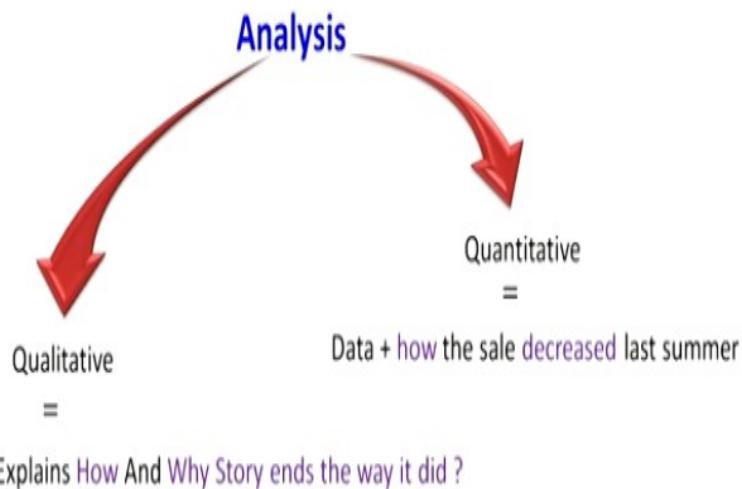
Okay, in the contrary the analytics is studying about what will happen in future and with the help of analytics. We can predict explore possible potential future events.

(Refer Slide Time: 12:25)



So the analytics is maybe qualitative or quantitative. For example in analytics if we say qualitative analytics. So it is the decision mostly based on the intuition. But if you say in quantitative where with the help of formulas with the help of algorithms will make the decisions.

(Refer Slide Time: 12:44)



So in the analysis data analysis also we can go for qualitative. We can explain how and why a story ends in that way it did? When we say in quantitative we can say, how the sales decreased the last summer. When I say as I am repeating, when you say analysis is something studying about what has happened in the past.

(Refer Slide Time 13:12)

Analysis ≠ Analytics

Data Analysis ≠ Data analytics

Business Analysis ≠ Business analytics

Okay, so it is not exactly analysis equal to analytics. Similarly when you say data analysis is different data analytics is different.

Similarly business analytics is different business analytics. When you say analytics is nothing but studying about the future events with the help of the past data.

(Refer Slide Time 13:34)

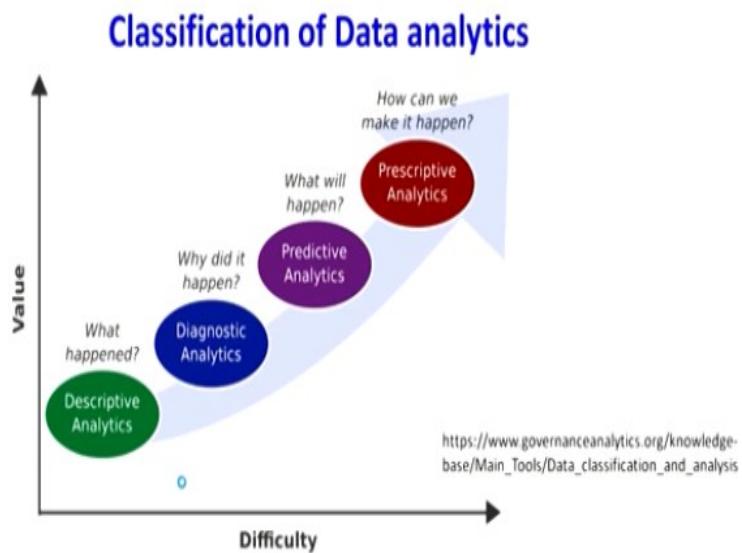
2.5 Classification of Data analytics

Based on the phase of workflow and the kind of analysis required, there are four major types of data analytics.

- Descriptive analytics
- Diagnostic analytics
- Predictive analytics
- Prescriptive analytics⁹

Next we will go for classification of data analytics, based on the phase of workflow and the kind of analysis required, there are four major types of data analytics. One is descriptive analytics, diagnostic analytics, predictive analytics and prescriptive analytics. We will see these four types of analytics in detail in coming classes:

(Refer Slide Time 13:57)



If we look at the difficulty and the kind of value which we can get from different types of analytics; this picture shows for example: when you see the descriptive analytics that will answer what happened? Diagnostic analytics, will help you to answer why did it happen? Predictive analytics will help you what will happen? Prescriptive analytics will help you to answer how can we make it happen? There is one context when you look at the level of difficulty you see that the descriptive analytics is the level of difficulty is very less.

And the contrary when you look at the prescriptive analytics the difficulty level is more and the value also, value in the sense business value which adds to you also more. so when there is a more difficulty there is a more value. Okay,

(Refer Slide Time 14:54)

Descriptive Analytics

- Descriptive Analytics, is the conventional form of Business Intelligence and data analysis
- It seeks to provide a depiction or “summary view” of facts and figures in an understandable format
- This either inform or prepare data for further analysis
- Descriptive analysis or statistics can summarize raw data and convert it into a form that can be easily understood by humans
- They can describe in detail about an event that has occurred in the past

Then we listen what is the descriptive analytics? Descriptive analytics is the conventional form of business intelligence or data analysis. It seeks to provide the depiction or summary view of facts and figures in an understandable format. These either inform or prepare data for further analysis. so descriptive analysis or we can say another way in statistics can summarize raw data and convert it into your form that can be easily understood by humans. They can describe in detail about an even that has occurred in the past. Okay,

(Refer Slide Time 15:40)

Example

A common example of Descriptive Analytics are company reports that simply provide a historic review like:

- Data Queries
- Reports
- Descriptive Statistics
- Data Visualization
- Data dashboard



Source: <https://www.linkedin.com/learning/478e9692-d13d-338f-907e-d76f0724d773>

Some of the examples of descriptive analytics is a common example of descriptive analytics are company reports that simply provide the historic review like: data queries, reports, descriptive statistics, data visualization and data dashboard. Okay,

(Refer Slide Time 16:00)

Diagnostic analytics

- Diagnostic Analytics is a form of advanced analytics which examines data or content to answer the question "Why did it happen?"
- Diagnostic analytical tools aid an analyst to dig deeper into an issue so that they can arrive at the source of a problem
- In a structured business environment, tools for both descriptive and diagnostic analytics go parallel

The next one will go to the diagnostic analytics. Diagnostic analytics is a form of advanced analytics which examines data or content to answer the question why did it happen? So we are diagnosing, suppose we are meeting a doctor for consulting, so he will try to understand why this has happened? Okay so that kind of analytics nothing but diagnostic analytics. So the diagnostic analytical tools aid and analyst to dig deeper into an issue.

So that, they can arrive at the source of the problem. So doctor also will identify you somebody has got some disease what was the sources of the problem. Similarly the diagnostic analytics also if something has happened for example the company's not performing well that diagnostic abilities will help you to identify what was the core reason for that. In a structured business environment tools for both descriptive and diagnostic analytics go parallel.

So when you look at the whether it is a prescriptive or diagnostic analytics, the tools, analytical tools which are using can be same only the purpose may be different.

(Refer Slide Time 17:09)

Example

- It uses techniques such as:

1. Data Discovery
2. Data Mining
3. Correlations

For example: data discovery, data mining, and correlations. These tools can be used for your prescriptive analytics also. Okay,

(Refer Slide Time 17:20)

Predictive analytics

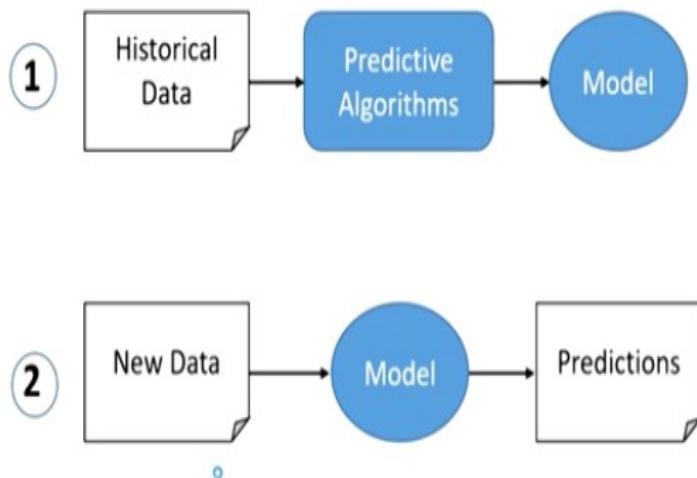
- Predictive analytics helps to forecast trends based on the current events
- Predicting the probability of an event happening in future or estimating the accurate time it will happen can all be determined with the help of predictive analytical models
- Many different but co-dependent variables are analysed to predict a trend
in this type of analysis

Now we will go for predictive analytics, predictive analytics helps to forecast trends based on the current events. When you say predicting obviously it says, that it is discussing about what will happen in future? Predicting the probability of an event happening in future are estimating accurate time it will happen can all be determined with the help of predictive analytical models. Many different but co-dependent variables are analysed to predict a trend in this type of analysis.

So in the predictive analytics one of the tool is the regression analysis. There may be some independent variables, some dependent variables, sometimes more dependent variable, more

than one dependent variable and how these variables are inter-related. So that kind of study is nothing but your predictive analytics.

(Refer Slide Time 18:11)



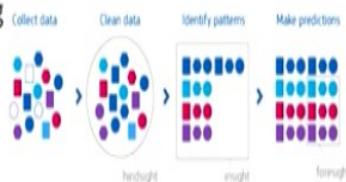
Source: <https://www.logianalytics.com/wp-content/uploads/2017/11/predictive-1.png>

When you look at this picture, you see that with the help of historical data by using different algorithm, predictive algorithms you can come with a model. Once the model is developed a new data can be fit into this model so we can get some predictions about the past events.

(Refer Slide Time 18:35)

Example

- Set of techniques that use model constructed from past data to predict the future or ascertain impact of one variable on another:
 1. Linear regression
 2. Time series analysis and forecasting
 3. Data mining



Source: <https://bigdata-madesimple.com/5-examples-predictive-analytics-travel-industry/>

Example is linear regression, time series analysis and forecasting and data mining. These are the techniques for predictive analytics.

(Refer Slide Time 18:46)

Prescriptive analytics

- Set of techniques to indicate the best course of action
- It tells what decision to make to optimize the outcome
- The goal of prescriptive analytics is to enable:
 1. Quality improvements
 2. Service enhancements
 3. Cost reductions and
 4. Increasing productivity

The last one is the prescriptive analytics. A set of techniques to indicate the best course of action. It tells what decision to make to optimize the outcome. The goal of prescriptive analytics is to enable: quality improvements, service enhancements, cost reductions and increasing productivity. Okay,

(Refer Slide Time 19:13)

Prescriptive analytics: Example

- Optimization Model
- Simulation
- Decision Analysis

In the prescriptive analytics, some of the tools which we can use is optimization models, simulation model, and decision analysis. These are the tools under prescriptive analytics.

(Refer Slide Time 19:27)

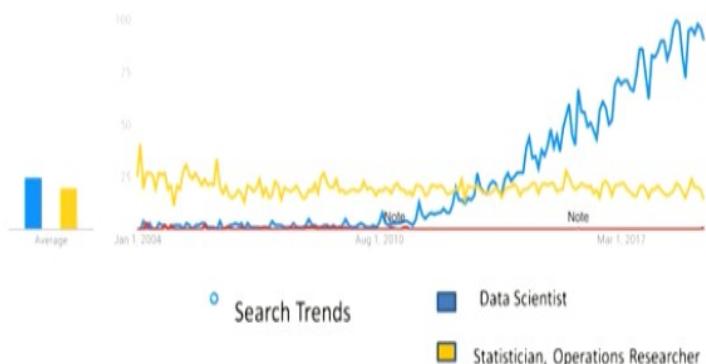
3. Explain why analytics is important

- Demand for Data Analytics
- Element of data Analytics

Next is we are going to see, why the analytics so important? In this section we will see what is happening the demand for data analytics and we look at the different elements of data analytics.

(Refer Slide Time 19:44)

3. Explain why analytics is important



This picture shows, Google Trends, this was up to 2017. See for example, the blue represents the data scientist; this orange represents the statistician operation researchers. You see the trend is it is increasing that means people are searching in the Google search engine the word data scientist more number of times. See the search count is increasing. That means there is a demand for that particular say job.

(Refer Slide Time 20:19)



You see, if you look at this is the newspaper clipping from Times of India. There are so many news are coming about data scientists and the future requirement of data scientists. You see the data scientist earning more than CA's and engineers. You can look at this link for further.

(Refer Slide Time 20:37)

3.1 Demand for Data Analytics

With companies across industries striving to bring their research and analysis (R&A) departments up to speed, the demand for qualified data scientists is rising.

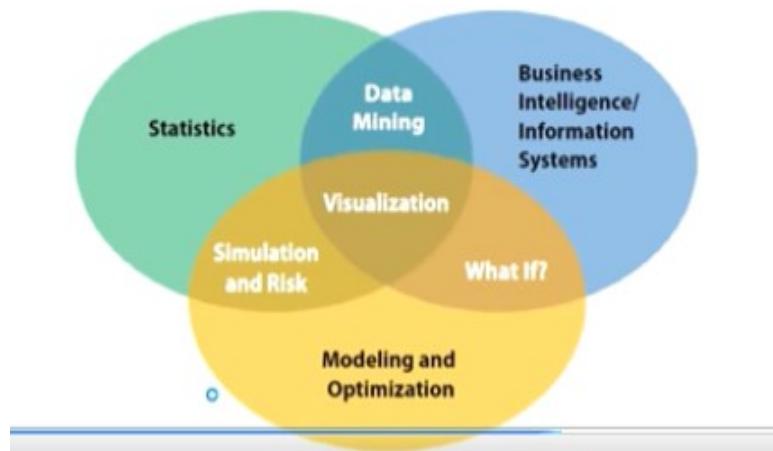
"India will face a demand-supply gap of 2,00,000 analytics professionals over the next three years. Even in the US, only 40 out of 100 positions for analytics professionals can be filled," said Rituparna Chakraborty, co-founder & senior VP of TeamLease Services.

http://timesofindia.indiatimes.com/articleshow/52171064.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst

And you see the demand for data analytics. This also newspaper clipping with companies across industries striving to bring their research and analysis department up to speed, the demand for qualified data scientist is rising. So there is an emerging field. so many companies are looking for the qualified data scientist. So if you take this course and end up the course that you may be qualified for getting into these companies.

(Refer Slide Time 21:07)

3.2 Element of data Analytics



Many times you see what is data analytics, Statistics, data mining, optimizations. These are students having different understanding on that. So when we say data analytics, there are different element one is statistics, next one is the business intelligence information systems, then modelling and optimizations, then simulation and risk. We can say if you are able to do what if analysis? That is nothing but sensitivity analysis, visualization, data mining. These are the components of data analytics and how these different domains are interrelated?

(Refer Slide Time 21:47)

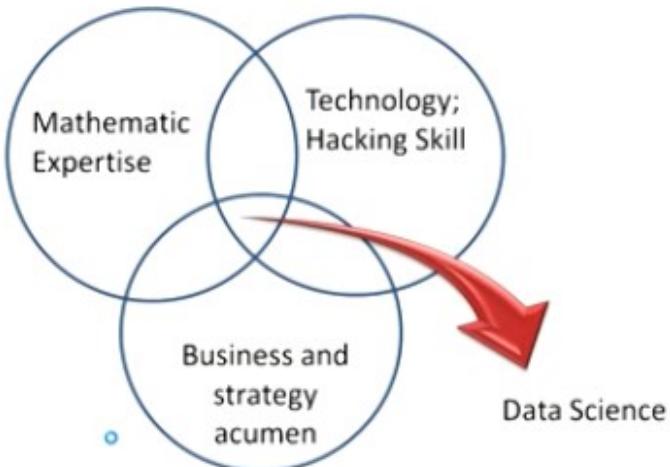
4. Data analyst and Data scientist

- The requisite skill set
- Difference between Data analyst and Data Scientist

Next we will see, what kind of skill set is required to become a data analyst? then we will see the small difference between data analyst and data scientist?

(Refer Slide Time 21:59)

4.1 The requisite skill set



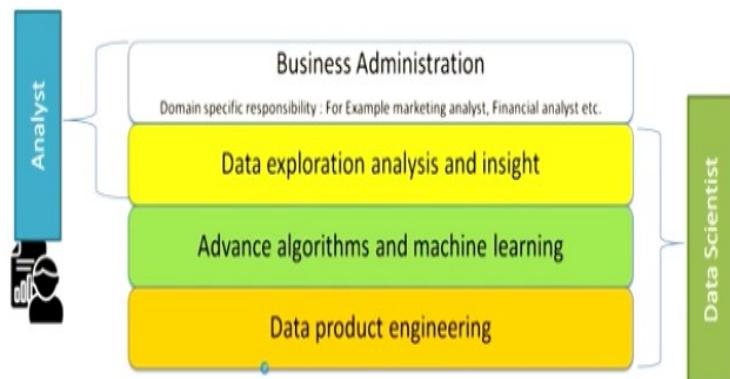
See to become a data analyst is the basic fundamental knowledge is you need to have knowledge of mathematics. Next you need to have the knowledge of technology is nothing but hacking skill. Hacking skills in the sense, if the data is given hacking is done and looked at the positive way. How to use the data to get more information? The next skill is business and strategy acumen; you should have the knowledge of the domain and knowledge of the business and you knew to the strategy equipment.

So these three skills are required for a good data scientist. It is very difficult to have a one person will have all these three skills that's why availability of good data analyst is becoming very difficult. Because somebody may be very good at mathematics but they may not have very good knowledge and business, some people may be very well at technology, technology in the sense information technology, they may not have good knowledge on the business knowledge.

So we need to have the combination of all these three skills otherwise the group of people some people from mathematics department or mathematics area, some people from computer science, some people from the domain knowledge. They were to work together to form a good data scientist team, so these forms data science.

(Refer Slide Time 23:31)

4.2 Difference between Data analyst and Data Scientist



Now what is the difference between data analysts and data scientists and the difference is what kind of role they are doing? For example; the role of your data analyst is, see in your business context, he may have the knowledge of business domain. For example; if he is good at doing analytics in the area of marketing, he can be called as a marketing analyst. If the person is from finance area, he can be called it as a finance analyst. So he is the analyst, data analyst.

But the role of data scientist is little bigger, because the data scientist need to have the knowledge of advanced algorithms and machine learning and able to come out with a data product. Which I told you in the previously, so the data scientist can come out with a data product. Okay,

(Refer Slide Time 24:30)

5. Why python?



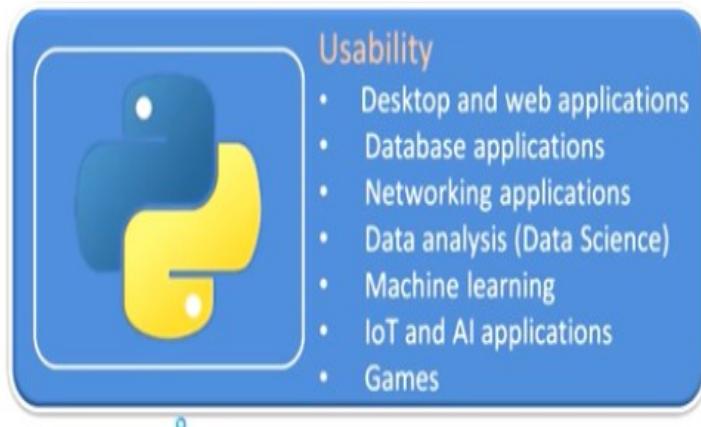
In this course we are going to use Python. In this in the next lecture, I will tell you the basic introduction about the Python. Here we will see why we are going to use the Python?. Because python is very simple and easy to learn. Most importantly it is a free software and open source. It uses interpreted, it is not the compiler. Suppose what do you my compiler and interpreter is you need a compiler to solve the whole program but interpreter need not be in that way.

It can solve, even you can interpret one sentence also, one line in the programming line also. it is dynamically typed, dynamically type in the sense in some other programs every time you have to declare the variable. What is the nature of the variable? Whether it is integer? Whether it is a float? But here you need not do. It is dynamically takes the value. it is extensible, extensible in the sense if you make a code in some other language that can be extended with the help of Python.

And can be embedded, embedded in the sense you have made some program in Python it can be embedded with the some other platforms and it has extensive library.

(Refer Slide Time 25:45)

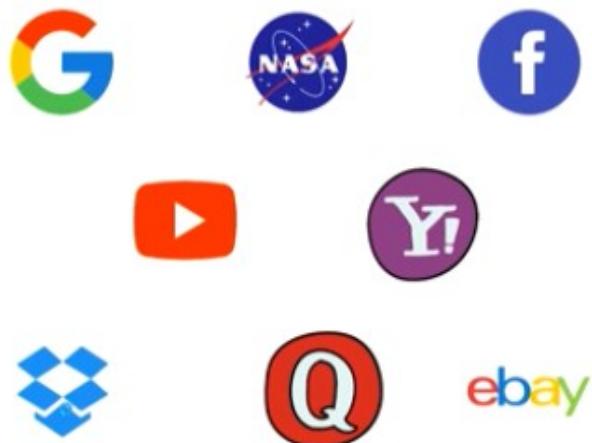
5. Why python?



The usability of Python is it is a desktop and web applications, it can be used for data applications, it can be used for networking applications, most importantly it can be used for data analyst, data science can be used for machine learning, it can be used for IoT Internet of Things and artificial intelligence applications and can be used for games.

(Refer Slide Time 26:05)

Companies using Python



Another reason for choosing Python is most of the companies, they use Python is a language in their company. Like for example; Google, Facebook, NASA, Yahoo and eBay. They use Python is a programming language.

(Refer Slide Time 26:23)

Why Jupyter NoteBook?



In this Python also we are going to use Jupyter notebook. In the next class I will explain you because it is the client-server application is edit code on web browser. It is easy in documentation, easy in demonstration and user friendly interface.

(Refer Slide Time 26:39)

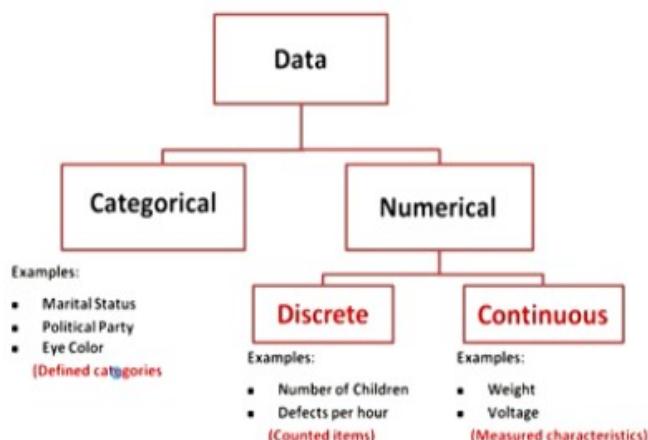
6. Explain the four different levels of Data

- Types of Variables
- Levels of Data Measurement
- Compare the four different levels of Data:
Nominal
Ordinal
Interval and
Ratio
- Usage Potential of Various Levels of Data
- Data Level, Operations, and Statistical Methods

This was the last session of this lecture; we will explain four different levels of the data. What is the type of variables? Levels of data measurement? Compare for different level of data: will say nominal, ordinal, interval and ratio. We will see that why and what is the usage of knowing this different level of data?

(Refer Slide Time 27:03)

6.1 Types of Variables



The one way for classifying the data is the categorical data, one is a numerical data. In categorical data; you see marital status, political party, and eye color. These are categorical data. Numerical data; it can be discrete or continuous. Discrete data may be a number of children and defects per hour. So this is the discrete data. In the continuous data may be weight and voltage. These are the example of continuous data.

So what is the difference between discrete and continuous is, you say a number of children you may say two children or three children 2.5 children was not possible but in continuous, if you look at between 0 & 1 the numbers are continuing there are infinite number of values that are there between 0 & 1. So it is a continuous variable.

(Refer Slide Time 27:56)

6.2 Levels of Data Measurement

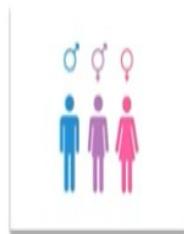
- Nominal – Lowest level of measurement
- Ordinal
- Interval
- Ratio – Highest level of measurement

Next will you see the different level of data measurement? Easily we have seen the classification of data. We classified as the categorical data and numerical data. There is another way of classification is, classifying into nominal data, ordinal data, interval data and ratio data.

(Refer Slide Time 28:14)

6.3.1 Nominal

- A **nominal scale** classifies data into distinct categories in which no ranking is implied
- Example : Gender, Marital Status



We will look at, what is a nominal data? Nominal scale classifies data into distinct categories in which no ranking is implied. The example of nominal data is gender, marital status. For

example; gender suppose you are conducting a questionnaire. Suppose you captured the gender male 0, female 1. This 0 1 represents just the gender. You cannot do any arithmetic operations with the help of the 0 & 1.

For example, you cannot find out the average, software will give you some number but there is no meaning for that. Similarly, marital status, whether it is married or unmarried. This is the example of nominal data.

(Refer Slide Time 29:01)

6.3.2 Ordinal scale

- An **ordinal scale** classifies data into distinct categories in which ranking is implied
- Example:
 - Product satisfaction → Satisfied, Neutral, Unsatisfied
 - Faculty rank → Professor, Associate Professor, Assistant Professor
 - Student Grades → A, B, C, D, F

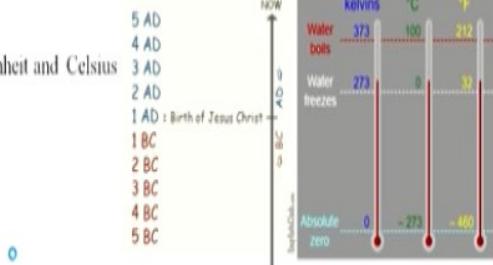
The next level of data is the ordinal scale. It classifies data into distinct categories in which the ranking is implied. Here the numbers are the ranked. For example; you may ask the customer to give a ranking about their level of satisfaction. For example, satisfied, neutral, unsatisfied. The faculty ranking, for example; professor, associate professor, assistant professor.

You see that their rank is followed for example 1 professor, 2 associate professor, 3 three professor. Student grades, A, B, C, D, E, F. These are ranking, because the numbers 1, 2, 3 represents the rank.

(Refer Slide Time 29:45)

6.3.3. Interval scale

- An **interval scale** is an ordered scale in which the difference between measurements is a meaningful quantity but the measurements do not have a true zero point.
- Example
 - Temperature in Fahrenheit and Celsius
 - Year



The next level of data is interval scale. The interval scale is ordered scale, in which the difference between measurements is a meaningful quantity but the measurement do not have to zero point. The example of interval scale is, for example year. Say now, this here is 2019, you can add and subtract something. You can add another five years, its 2024 or you can subtract another nine years, its 2010.

But you cannot multiply, if you multiply that number for example 2019 and 2020 you will end up with the big number there is no meaning for that. Because, there is no meaning for zero. Another example of interval scale is your Fahrenheit temperature. For example, in the Fahrenheit scale, the zero represents freezing point but it is not the absence of the seat but absence of the temperature but at the same time in the Kelvin for example minus 273 it is absence of heat. So Kelvin will be the some other scale. That you will see the next one,

(Refer Slide Time 30:52)

6.3.4 Ratio scale

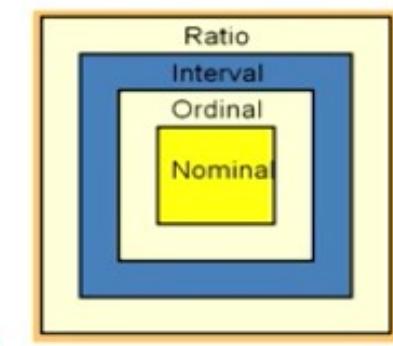
- A **ratio scale** is an ordered scale in which the difference between the measurements is a meaningful quantity and the measurements have a true zero point.
- Example
 - Weight
 - Age
 - Salary

The ratio scale is the ordered scale in which the difference between the measurements is a meaningful quantity and the measurements have the true zero point. Weight, age, salary and the Kelvin temperature comes under ratio scale. Because 0 Kelvin that means the absence of the heat. So in the ratio scale, he can do all kinds of arithmetic operation. For example the nominal, you cannot do any arithmetic operation. In ordinal you cannot do in arithmetic operation.

In the interval you can add and subtract but you cannot multiply. But in the ratio data, you can do all kinds of arithmetic operations. You can add. You can subtract, you can multiply, and you can divide.

(Refer Slide Time 31:35)

6.4 Usage Potential of Various Levels of Data



You see the usage potential various level of data. For example the usage potential of nominal data is not that much. The next one is ordinal; next one is interval, next one ratio. So the ratio data is having the highest to use its potential. The nominal data is having the least usage potential.

(Refer Slide Time 31:56)

6.5 Impact of choice of measurement scale

Data Level	Meaningful Operations	Statistical Methods
Nominal	Classifying and Counting	Nonparametric
Ordinal	All of the above plus Ranking	Nonparametric
Interval	All of the above plus Addition, Subtraction	Parametric
Ratio	All of the above plus multiplication and division	Parametric

This is more important, why we have to still know the different types of data. Because this types of data is helping to choose the right analytical tools for doing analysis. For example; if the data is the nominal data. You can do only nonparametric tests. For example the data is ordinal, here also you can do only nonparametric test. But if the data is interval, you can do parametric test. You see that interval; you can do all above plus addition and subtraction.

In the ratio, if you can do all of the above plus multiplication and division and statistical methods. You can go for parametric methods. So the purpose of classifying the data into nominal, ordinal, interval, ratio is to choose the right analytical tools with it whether it is a parametric or non parametric. The other reason is sometime for if we want to do a non parametric analysis that is used only for nominal data.

Sometime the students they will, the data may be nominal but they may go for a parametric test that, should not be done. That is the purpose of knowing what kind of, what is the nature of this data. So in this class we have seen the introduction for data analytics. We have seen the importance of data analytics. We have seen the classification of data analytics. Then we can we have seen what is the analytics and analyst and we have seen different types of data.

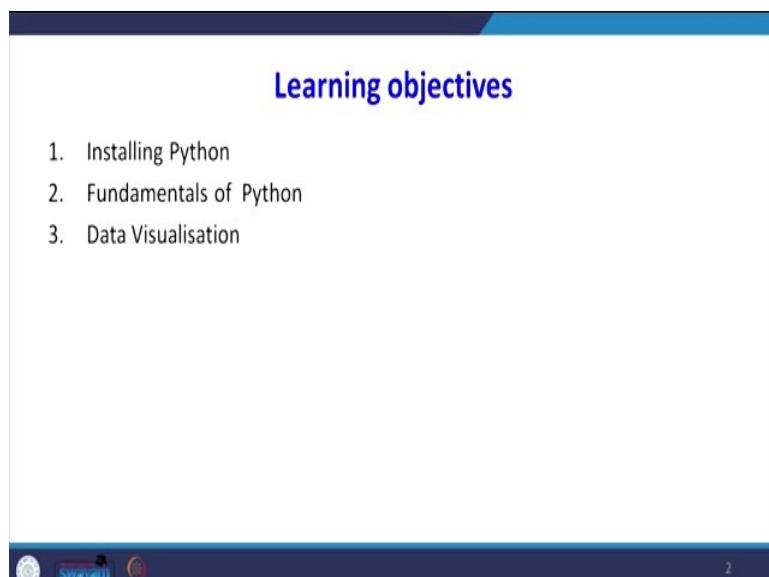
The next class we will learn about what is Python? How to install the Python and what kind of descriptive analysis we can do with the help of Python? So the next class will meet you with another lecture. Thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology, Roorkee

Lecture No 2
Python fundamentals -1

Good morning students, in the last class, that was the introduction class, we have seen the importance of data analytics and we have seen certain classification of data analytics. This is my second lecture that is Python fundamentals because we are going to use this Python. In this lecture I have 3 objectives.

(Refer Slide Time: 00:50)



The slide has a blue header bar. The main title 'Learning objectives' is in blue. Below it is a list of three items:

1. Installing Python
2. Fundamentals of Python
3. Data Visualisation

At the bottom of the slide, there is a dark footer bar with some icons and the number '2'.

One is I will tell you how to install Python second one I will see some fundamentals of the Python, third one some data visualization. In the data visualization I am going to give only theory in this class. The next class we are going to use Python and we are going to take some sample data and we have to visualize the data using Python software.

(Refer Slide Time: 01:13)

Python Installation Process

Installation Process –

Step 1: Type <https://www.anaconda.com> at the address bar of web browser.

Step 2: Click on download button

Step 3: Download python 3.7 version for windows OS

Step 4: Double click on file to run the application

Step 5: Follow the instructions until completion of installation process



As I told you the 1st one is how to install this Python. There are 5 steps is there. Step 1, we are going to see in detail in coming slides. In step 1 we are going to visit this website www.anaconda.com at the address bar of the web browser 2nd one we are going to click on download button 3rd one will download python 3.7 version for Windows operating system. Then we will double click that is a 4th step we will double click on file to run the application.

The 5th one will follow the instruction until the completion of installation process. What I have done I have taken the screenshot of all these, all the 5 steps while installing the laptop. I am going to show each steps in the form of screenshot.

(Refer Slide Time: 02:03)

Python Installation Process

Installation Process –

Step 1: Type <https://www.anaconda.com> at the address bar of web browser.



The 1st one is type this www.anaconda.com at the address bar of the web browser.

(Refer Slide Time: 02:13)

Python Installation Process

Step 2: Click on download button

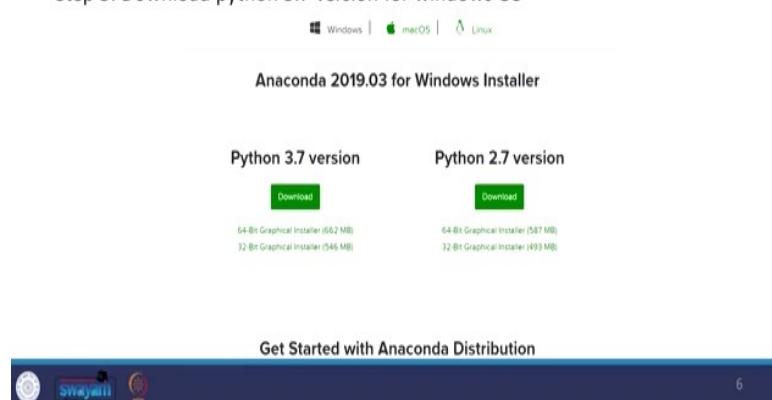


2nd one is once you typed it you can see this screen, here you see this location you can see here. This location there is a download option. When you click that you see that the left side also I have rounded there is a download option you download it.

(Refer Slide Time: 02:36)

Python Installation Process

Step 3: Download python 3.7 version for windows OS

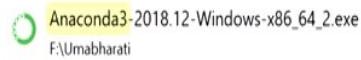


In the 3rd step is there are two versions of python, python 3.7 and python 2.7. In these courses we are going to use the latest version that is the Python 3.7.

(Refer Slide Time: 02:46)

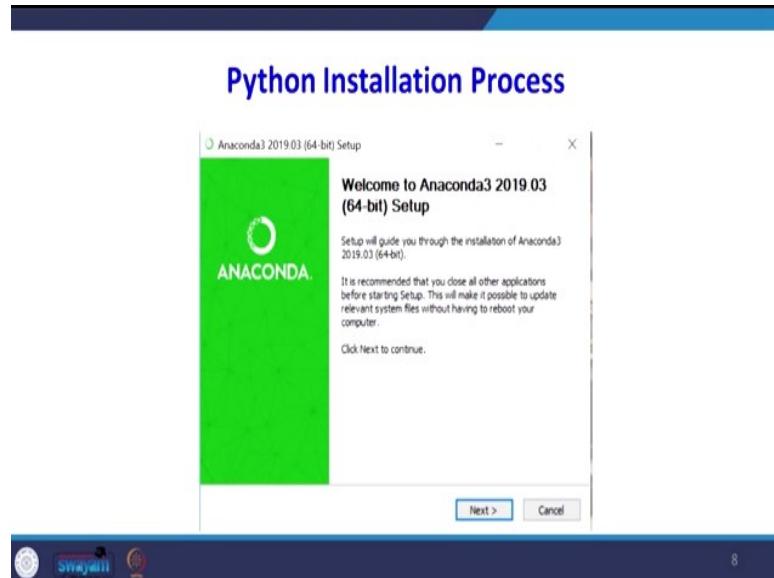
Python Installation Process

Step 4: Double click on file to run the application



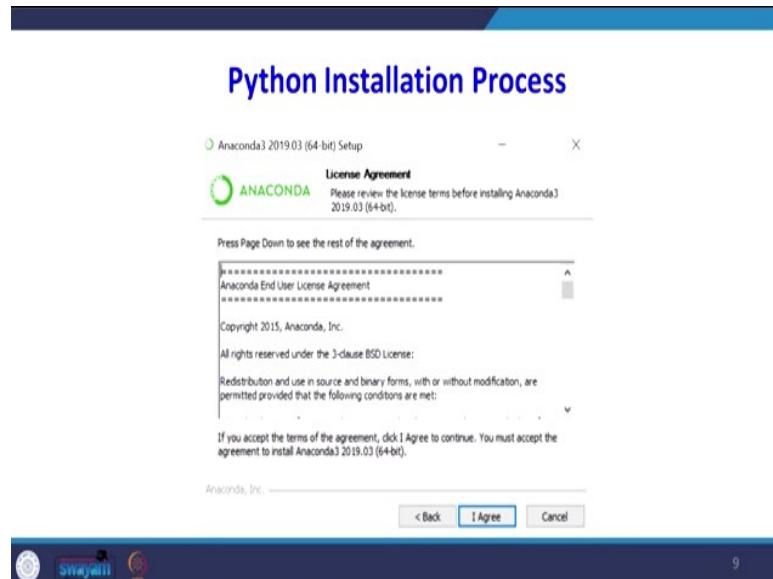
In the 4th step double click on file to run the application it will get downloaded. when you double click for example; I have stored this anaconda in F folder.

(Refer Slide Time: 02:59)



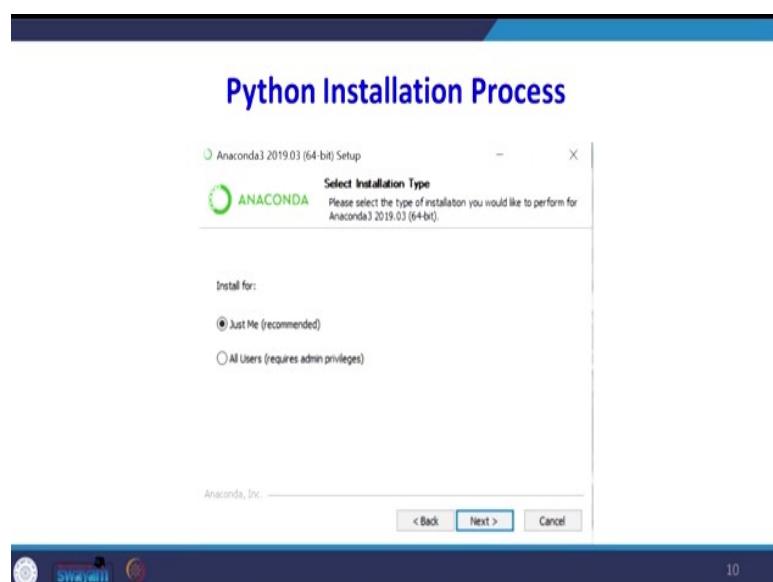
Step 5 is just to keep on click Next

(Refer Slide Time: 03:02)



You have to agree for their agreements, terms and conditions.

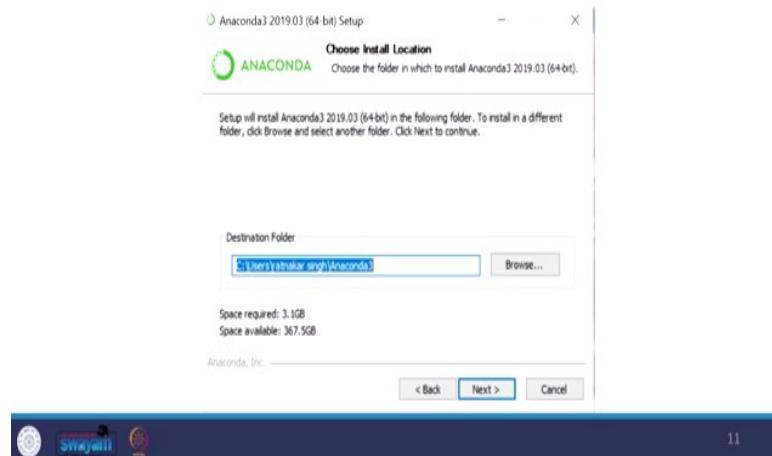
(Refer Slide Time: 03:06)



Next you select, just me recommended click Next.

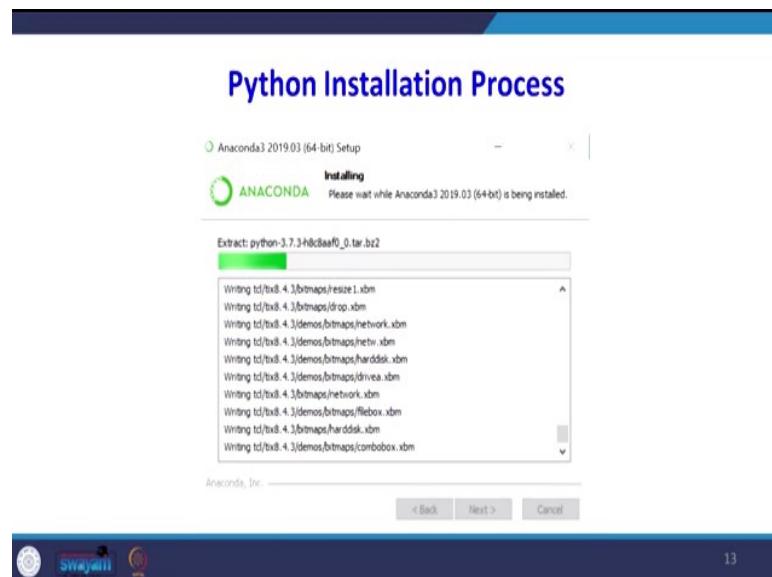
(Refer Slide Time: 03:10)

Python Installation Process



Then it is installed in C drive click Next.

(Refer Slide Time: 03:15)



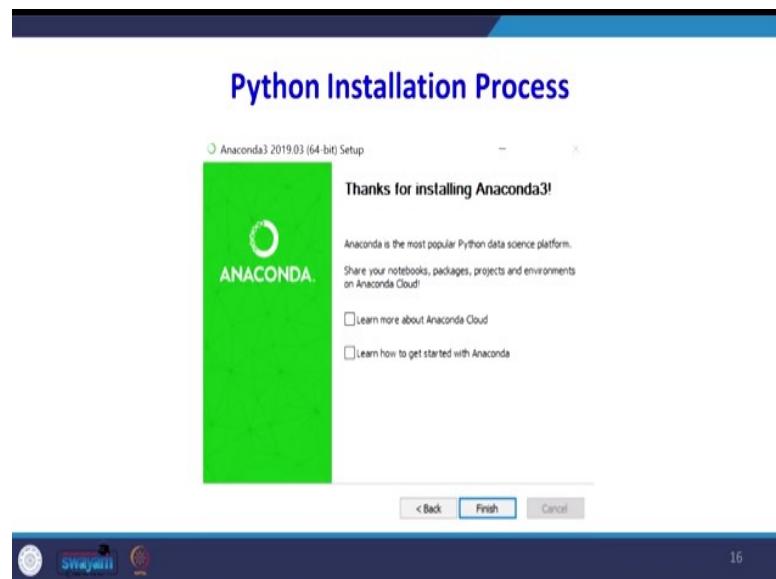
Then install, Installation process is started, then installation is completed.

(Refer Slide Time: 03:25)



Again click Next

(Refer Slide Time: 03:29)



Then click and finish Ok

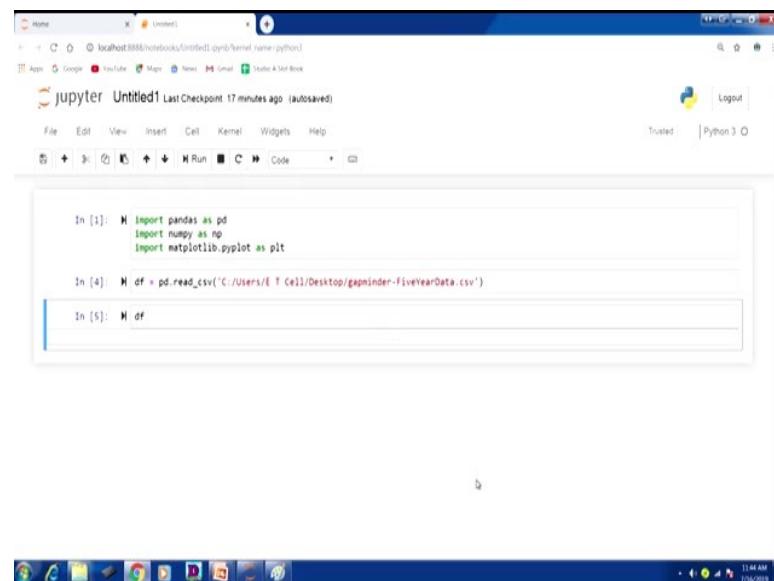
(Refer Slide Time: 03:34)

Why Jupyter NoteBook?



Now we have installed the anaconda. So I will explain you how to open Jupyter notebook. I will switch the screen

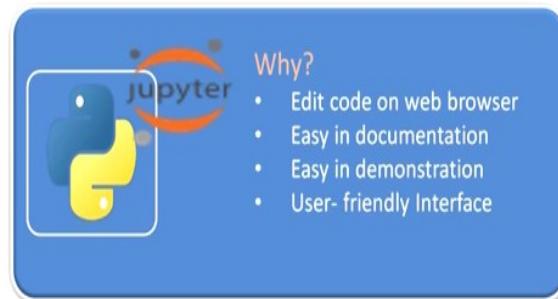
(Refer Slide Time: 03:46)



Yeah, this is the screen. The initially there are you see, I will see what is this some box it is showing in blue color sometimes it will show in green color that I will show you later. So this is the Jupyter notebook look like.

(Refer Slide Time: 04:04)

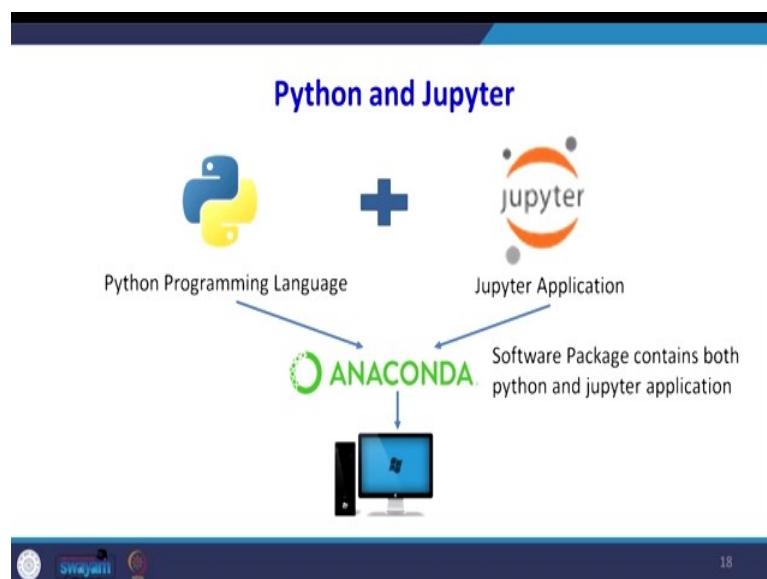
Why Jupyter NoteBook?



17

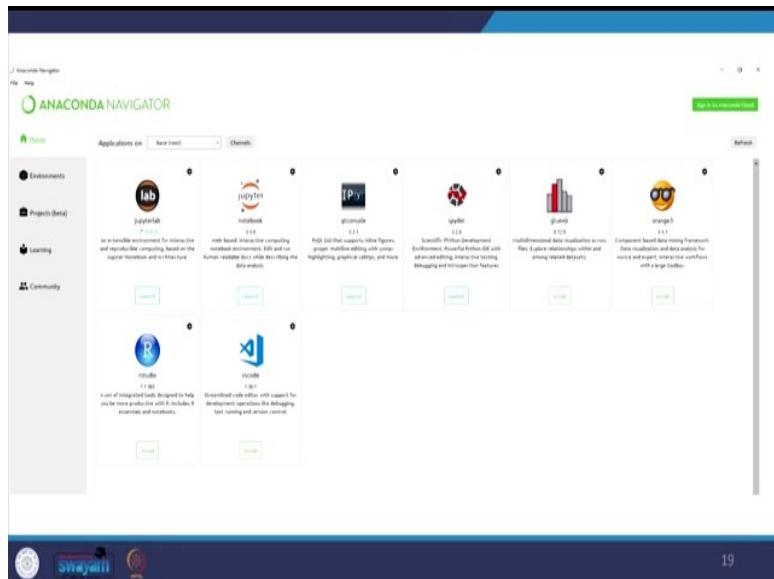
The next one is why there are some more interfaces there for using Python. There is a spider is there Jupyter is it but we prefer Jupyter for some reasons because it is edit code on web browser, it is easy in documentation, it is easy in demonstration and it is user-friendly interface. That was the reason we are using Jupyter it is not necessary if you already you are comfortable in some other interface you can continue with that.

(Refer Slide Time: 04:34)



See that in anaconda it consists of two software one is python that is on the left hand side the another side the right hand side the Jupyter applications these are combined together and kept in the Anaconda software package.

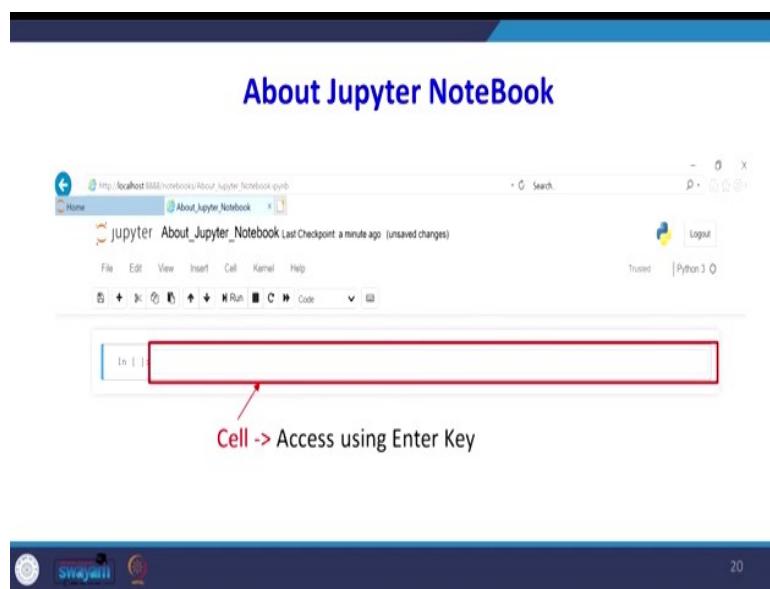
(Refer Slide Time: 04:49)



19

When you from the start when you type Jupyter you will get this screen

(Refer Slide Time: 04:58)



20

Then when you click launch you will get this one. So now from the start I am going to explain how to start this Python jupyter notebook.

(Video Starts: 05:08)

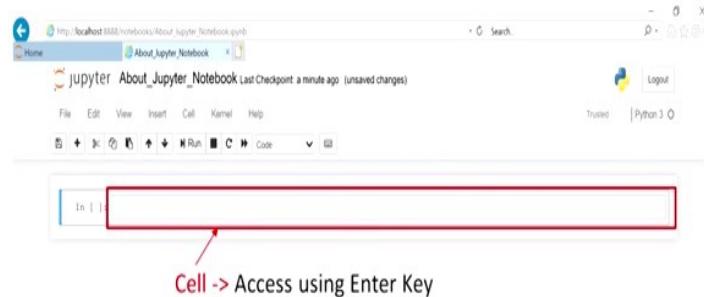
You have to type Jupiter and Jupiter notebook . When you click it you will get this one. Suppose if you want to type in a new go new Python 3. Yeah? Here there is a Jupyter there is a it is coming untitled 2 there you can change the name. You give the name as introduction to Python, introduction Okay?

(Video Ends: 05:40)

(Refer Slide Time: 05:40)



About Jupyter NoteBook

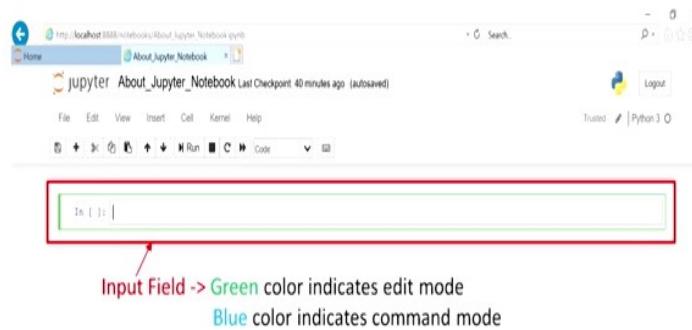


You see there is a box is appearing this is called cell I have made it in the red color, it is a cell it can be the cell can be accessed using Enter key.

(Refer Slide Time: 05:51)



About Jupyter NoteBook

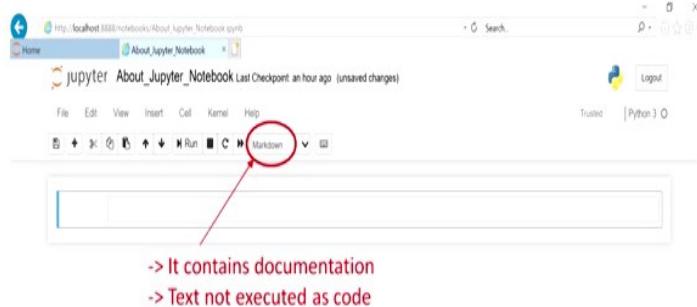


You see sometime that box will look like a green color, Green color indicates it is in edit mode sometime the box will look like in blue color.

(Refer Slide Time: 06:01)



About Jupyter NoteBook



The blue color indicates it is a command mode. See when you go to below the help there is a file name is called mark down .There if you type something then you select mark down that is used for making documentation. So it contains documentation, here text not executed as a code it is only for our understanding purpose.

(Refer Slide Time: 06:22)



About Jupyter Notebook

- Command mode allow to edit notebook as whole
- To close edit mode (Press Escape key)
- Execution (Three ways)
 - Ctrl +Enter (Output field can not be modified)
 - Shift +Enter (Output field is modified)
 - Run button on Jupyter interface
- Comment line is written preceding with # symbol.



Okay? Now about the Jupyter Notebook Command mode allowed editing notebook as a whole. To close edit mode press Escape key. Execution can be done in three ways you can simultaneously we can press Ctrl+Enter. So what will happen when you press Ctrl+Enter output field cannot be modified, another way is to press Shift+Enter output field is modified. Then there is a third way is there is a run button on the jupyter interface.

That you can directly you can click that. Then your code will get executed command line is written proceeding with # tag symbol. So when you want to make some understanding on your program you can use the # symbol, so that will not be executed.

(Refer Slide Time: 07:17)

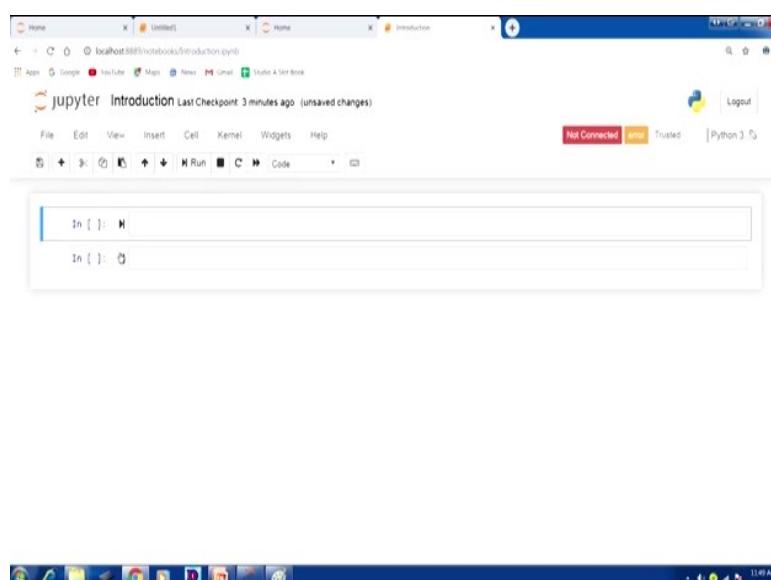
The screenshot shows a presentation slide with a blue header and footer. The title 'About Jupyter Notebook' is at the top. Below it is a bulleted list of shortcut keys:

- Important shortcut keys
 - A -> To create cell above
 - B -> To create cell below
 - D + D -> For deleting cell
 - M -> For markdown cell
 - Y -> For code cell

The footer contains icons for 'SWAYAM' and 'SWAYAM', the number '24', and a small logo.

That you only for your understanding purpose but there are about the Jupyter notebook important shortcut keys. When you press A that is used to create a cell above when you press B that is to create a cell below when you press D+ D for deleting cell. When you press M that will made a say mark down cell, when you press Y that is for coding cell.

(Refer Slide Time: 07:46)



For example; when I am entering B

(Refer Slide Time: 07:54)

Fundamentals of Python

- Loading a simple delimited data file
- Counting how many rows and columns were loaded
- Determining which type of data was loaded
- Looking at different parts of the data by subsetting rows and columns

25

We will go to the next one fundamentals of Python and you see loading here .What we are going to see in coming slides. Loading a simple delimited data file counting how many rows and columns were loaded and determining which type of data was loaded. Then looking at different parts of data by subsetting rows and columns because these activities are more important because once we loaded a data that may have n number of cells n number of rows, column and rows.

Sometime we need to do some operation using only few rows are few cells .You should know how from a big data file how to use only a particular row or how to use a particular column. Sometimes we can have a collection of rows also, collection of columns also for doing our specific operations.

(Refer Slide Time: 08:49)

Pandas for Everyone

Python Data Analysis

Daniel Y. Chen

♦ Addison-Wesley

Boston • Columbus • Indianapolis • New York • San Francisco • Amsterdam
Cape Town
Dubai • London • Madrid • Milan • Munich • Paris • Montreal • Toronto •
Delhi • Mexico City
Sao Paulo • Sydney • Hong Kong • Seoul • Singapore • Taipei • Tokyo

26

This was the reference book which I am following for this course and the book name is Pandas for everyone especially for this lecture. It is the professor Daniel Y. Chen he is the author of this book.

(Refer Slide Time: 09:04)

Loading a simple delimited data file

```
In [1]: import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
  
In [23]: df = pandas.read_csv('F:/2019-20/IPTEL/2 Introduction to Python/data/gapminder-FiveYearData.csv')  
  
In [28]: df
```

Data Source: www.github.com/jennybc/gapminder.



Now we are going to learn how to load a simple delimited data file. This is the fundamental because before doing data analysis the first step is how to load the data into the Python. For that purpose we are going to import some basic libraries one is pandas numpy another is matplotlib.pyplot as plt. So, first we are going to import these three basic library .Then we are going to load the data. The data sources it is taken from www.github.com/jennybc/gapminder.

So I have downloaded this data set already I am going to tell you how to load the data set in to the python. Before that I am going to open that excel file, I am going to show what is the column? What is the row open the excel file?

(Video Starts 10:07)

When you look at this I am reading the column see that there is a country, year, population, continent, life expectancy that is given as the short name life exp then gdp per capita. So in rows there are, how many rows is there I will tell you how many rows is there I am coming down and this is a this is csv file format. How many rows are there? There are 1705.

The last row is Zimbabwe Right? Please look at the data Zimbabwe year 2007. I think it is a population, continent, life expectancy this is a per capita income. Okay? Now this data, this csv file I am going to import into the Python. You see that I am going to call this data set df.

`df= pd` because `pd` is the short-form of pandas, Pandas nothing but the panel data `Pandas.read_csv`.

Why I am using csv because the csv file is I am going to read it. The location of the file given the path of that file you can directly copy that path but one thing you have to note it down. See, C: this will be this should be \ because when you copy that path directly. Generally you will get here / but you have to change it. So I changed it back C: / users / ET cell / desktop / gapminder-five year data.csv.

Look at their it should be in the code. Now I am going to read the df, Yes? once I read it you see that, the row is starting from 0 . That is a very important. It is a 0 indexing 0, 1, 2, 3, 4, 5, 6, 7, 8 I am able to see whatever I have seen in the csv file just a few minutes before. You see I am able to see the country, year, population, continent, life expectancy and gdp per capita. Okay? What I how I have read it `pd.read_csv`

Suppose I have installed, I have loaded that data I want to say what are the headings of that file. Heading means what are the columns. For that there is, in Python there is a two type print and open the parentheses `df.head` when you execute this one you will get 1st 5 rows that means 0, 1, 2, 3, 4. So that means when you execute this one you can see 1st 5 rows from the data set Yes? You are able to see that, Okay?

I will go to the next command, suppose I want to know the size of that file that is I want to know how many rows and how many column is there. For that there is a command called the shape. So print `df.shape`, `df` is they were finally because we outer loading that csv file we have named in the variable called `df`. Okay? So when you type print `df.shape` then we will come to know how many rows are there. How many columns are there?

So, I am typing print `df shape`. One more thing you should not type this parenthesis because it is the shape is without parentheses. So I am going to remove this parenthesis again I got to run it. Yes, it is showing how many rows? How many columns? Okay? We will go to the next one now I want to know how many column names? What are the column names? So if I type print `df.columns` Right?

Here, please note that here also there is no parenthesis if I type print df.columns. This was the output which I copied see what is output disappearing country, year, population, continent, life expectancy, gdp per capita, data type is object; I will show you how this comment is running. Type print df., yes you are able to get this way. So what the students what you have to do while looking at the video you have to open your laptop you have to type this command.

Then you have to see you can verify the answer. Okay? The next command is to get the data type of each column; you have to type this command print df.dtypes. That will give you the summary of the all data set and what is the nature of the data. We will see that how it is appearing. So, I am going to type print df.dtypes. Now you we will see the data type of each column. For that you have to use this command dtypes.

So print df.dtypes, this is the output which you will get it. I will show you in the Python, first we will look at what is the subject output you see countries object, year is an integer, population pop it is a variable that is in the float. Float means there is a decimal continent it is an object that means a character, life expectancy it is a float that means you are going to get that value in decimals.

Similarly, gdp per cap that also going to get in the decimals then data type is object now we will go to the Jupiter. We run this command so you see that you see line number 8. Print df.dtypes you are getting whatever it was there in this or whatever I have shown in the slide is there. Say country object, year integer, population float type of data and so on.

(Video Ends 17:10)

(Refer Slide Time: 17:11)

Pandas Types Versus Python Types

Pandas Type	Python Type	Description
object	string	Most common data type
int64	int	Whole numbers
float64	float	Numbers with decimals
datetime64	datetime	datetime is found in the Python standard library (i.e., it is not loaded by default and needs to be imported)

33

This is a classification of types of data in the perspective of pandas, in the perspective of Python. See when they say string it is a most common data type it is a character. When I say ‘Int’ it is a whole number integers. It is a float number with the decimals. Date, time is that is to represent the data it is not loaded by default that need to be imported. Whenever it is a requirement is there that we will see

(Refer Slide Time: 17:38)

get more information about data

```
print(df.info())
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1704 entries, 0 to 1703
Data columns (total 6 columns):
country      1704 non-null object
continent    1704 non-null object
year         1704 non-null int64
lifeExp       1704 non-null float64
pop          1704 non-null int64
gdpPercap    1704 non-null float64
dtypes: float64(2), int64(2), object(2)
memory usage: 80.0+ KB
None
```

34

That one more command is to get more information about the data. So you type df.info you will get the full details about each columns. We will do that one.

(Video starts: 17:53)

Look at this when it print df.info so I am getting data columns there were 6 columns country there are 1704 rows is there Non null object that means all the data is there is no missing values. Similarly year 1704 rows is there, Non-null that is an integer, Non null means, that all

the values are filled. There is no missing cell so population, float, continent object, life exp float, gdp per capita float, memory usage is this much.

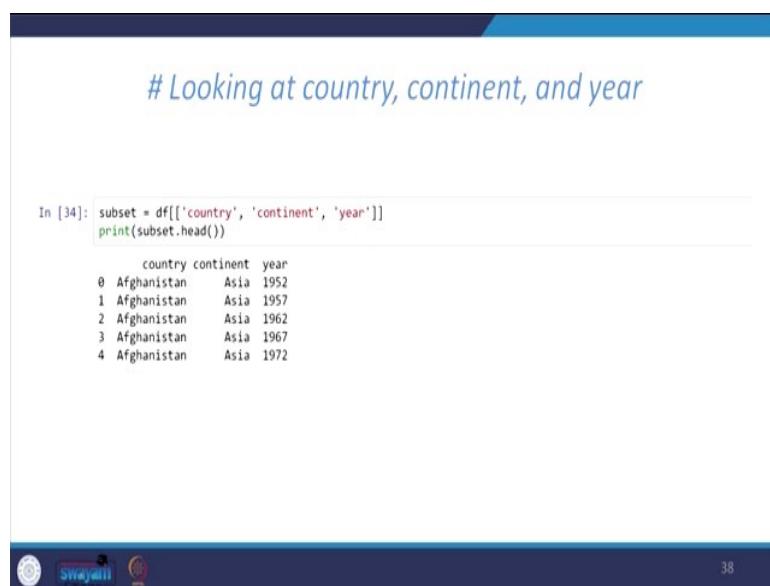
Suppose, there is a big data file is there we want to see the specific rows are specific columns. How to do that? Now get the country column and save it to its own variable. So country if you look at the data which I will show initially countries one of the column. So I want to pick up only that country column I am going to save it. I am going to give the name for that a country_df= df you see that you have to open the square bracket, Square bracket within quote.

Suppose, in the country column I want to see 1st 5 rows Okay? You type print, open parenthesis country_df.head that shows 1st 5 rows and see that now from the full data we have fetched only the country column. That we have seen there are 1st 5 rows, that is 0th row is a, 0, 1, 2, 3, 4, 5 to 5 rows we are able to see when you from the big file. Suppose there may be requirement you need to see what are the last five observations for that purpose.

You type print country_df.tail, then you can see from the bottom we can see last 5 rows. You will see how it is appearing, Yes? So what is it we are able to see last 5 rows from the country, country_df file.

(Video ends: 21:15)

(Refer Slide Time: 21:15)



```
# Looking at country, continent, and year

In [34]: subset = df[['country', 'continent', 'year']]
print(subset.head())

   country continent  year
0  Afghanistan     Asia  1952
1  Afghanistan     Asia  1957
2  Afghanistan     Asia  1962
3  Afghanistan     Asia  1967
4  Afghanistan     Asia  1972
```

There may be requirement you need to see more than one column at a time. So I am going to save in the form of another file name that is called a subset, Subset=df. You see there is a double square bracket so I want to switch the country columns, continent columns and year columns. Then I going to see what are the heading that means I want to see what does the 5 rows of these subsets so we will go to they go to Python.

(Video starts: 17:46)

I am going to call it a subset continent. Suppose I want to see the 1st 5 rows of this file called a subset. Data set called subset. You see that I am able to fetch 3 columns at a time that is on the country, continent and year. The same way we from the subsets file I want to see last 5 rows so print subset.tail. Let us see what we are getting we will get this output Yes? You see that there are 3 columns.

There were the last 5 rows from the button. So far we looking at different columns now we want to subset rows by index label there is one command called loc. So first we look at the file initial file that is a print df.head. Next you see that I want to locate the 0th row so for that purpose, print df.loc see it is a square bracket you type 0 because if you suppose we want to know the 1st row i out to enter because I would enter 0 because Python counts from 0, so print df.loc 0 that will show the 1st row.

You see 0th row access country Afghanistan year 1952 population is this much continent is Asia. Suppose I want to access this 1st row that means 0th row, Yeah? 0th row you can verify 0th row is the country Afghanistan, year 1952 this is a way to access a particular row. Dear students whatever comments which I am typing that I will be given to you when you take this course you can practice yourself.

You need not bother about in case we are not getting at this stage this all the commands all the codes will be given to you .You can practice on yourself. Suppose I want to get the 100th row how to access from the file df? I want to look at 100th row so you type print df.loc 99. You can exactly access in 100th row what is the element is there? Suppose I want to access 100th row df, 100th row is the country Bangladesh, year 1967, population this is.

This is the way to access different rows for our calculation purpose. So far we have seen how to load csv file into the Python, we have seen some basic commands.

(Video Ends: 25:21)

We have seen how to know the size of the file then we have seen how to access a particular row and also we have seen how to subset from the given big file? How to subset different small data file? So that can be used for our further analysis. So the next class we will see how to access different columns that will continue in the next lecture. Thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology, Roorkee

Lecture No 3
Python fundamentals

Okay? We will continue our lecture. How to access different rows and columns, because, it has very important applications.

(Refer Slide Time 00:33)

Looking at Columns, Rows, and Cells

- **Subset Rows by Index Label: loc**

```
In [36]: print(df.head())
```

	country	year	pop	continent	lifeExp	gdpPerCap
0	Afghanistan	1952	8425333.0	Asia	28.801	779.445314
1	Afghanistan	1957	9240934.0	Asia	30.332	820.853030
2	Afghanistan	1962	10267083.0	Asia	31.997	853.100710
3	Afghanistan	1967	11537966.0	Asia	34.020	836.197138
4	Afghanistan	1972	13079460.0	Asia	36.088	739.981106

When the data file is very big sometimes you need to access only some rows or some columns for your calculation purpose. That we will learn how to access a particular rows or particular columns there is looking at columns, rows and cells. When you look at this see print df.head when I use this command and getting there are different. For example; the first column says 0, 1, 2, 3, 4, country, year, population, continent, life expectations, gdppercapita.

(Refer Slide Time: 01:04)

get the first row

- Python counts from 0

```
In [37]: print(df.loc[0])
```

```
country      Afghanistan
year          1952
pop           8.42533e+06
continent     Asia
lifeExp        28.801
gdpPercap     779.445
Name: 0, dtype: object
```

3

Suppose I want to get the first row as we know that the Python counts from 0. If you want to know the first row you type a print df.loc, it is a location in square bracket 0. Will do that you will get the details which are there in the first row.

(Refer Slide Time: 01:26)

- # get the 100th row
Python counts from 0

```
In [38]: print(df.loc[99])
```

```
country      Bangladesh
year          1967
pop           6.28219e+07
continent     Asia
lifeExp        43.453
gdpPercap     721.186
Name: 99, dtype: object
```

4

So first if I want to know hundredth row so printed df.loc 99. We knew that python count from zero. If I want to know 100th row you have to type 99. So it should be in Square bracket you can see the details in the 100th row.

(Refer Slide Time: 01:42)

- get the last row

```
In [39]: print(df.tail(n=1))
   country  year      pop continent  lifeExp  gdpPerCap
1703  Zimbabwe  2007  12311143.0    Africa  43.487  469.709298
```



5

Suppose we want to know the last row in the data set. So print df.tail n equal to 1. If you type n equal to -1, it will not work, that we will see why if you want to know the last row simply type to df.tail n equal to 1, you will get to know that what is the last two, So we will see that.

(Video Starts: 02:01)

Now we are going to use this command to see the last row, that is a detail about the last rows. Now we can subset a multiple rows at a time. For example; there will be requirement we have to select 100th row, 1st row 100th rows and 1000th rows. For that purpose you type this command print df.loc. You see there are two square brackets 0, 99, 999, you will see what output where getting. So type print df.loc. Yes, so we are able to see the 1st row, 100th row, 1000th row.

There is another way we can subset rows by row number by using this command iloc. Previously loc, now we are going to use iloc. Suppose for type I want to get the 2nd row, if I type print df.iloc 1. I will get the details about the 2nd row. Okay? Yeah, this is a detail about the 2nd row. Suppose I want to know 100th row by using iloc command so go there. Yes? That is the details about the 100th row. You see that if I want to access the last row by using iloc command.

So you can directly type print df.iloc in squared bracket - 1. So that will be the details of the last row. So what you can do we can open our Excel file you can verify what was the title, the last row and soon.

(Video Ends: 04:27)

(Refer Slide Time: 04:27)

With iloc, we can pass in the -1 to get the last row—something we couldn't do with loc.

See then important note here with iloc command. We can pass in the -1 to get the last row, but same thing that we could not do with loc. That is the difference between loc command and iloc command.

(Refer Slide Time: 04:42)

- # get the first, 100th, and 1000th rows

In [44]: `print(df.iloc[[0, 99, 999]])`

	country	year	pop	continent	lifeExp	gdpPerCap
0	Afghanistan	1952	8425333.0	Asia	28.801	779.445314
99	Bangladesh	1967	62821884.0	Asia	43.453	721.186086
999	Mongolia	1967	1149500.0	Asia	51.253	1226.041130

Suppose we want to get the first 100th and 1000th rows, using iloc command. So we are going to type this print df.iloc 0, 99, 999. Let us see what answer we are getting.

(Video Starts: 04:58)

Yes? See, we have getting 1st, 100th and 1000th row.

(Video Ends: 05:21)

(Refer Slide Time: 05:21)

Subsetting Columns

- The Python slicing syntax uses a colon, :
- If we have just a colon, the attribute refers to everything.
- So, if we just want to get the first column using the loc or iloc syntax, we can write something like df.loc[:, [columns]] to subset the column(s).

12

So far we are seeing subsetting rows. Now we will see subsetting columns, the Python slicing syntax used a colon, colon represents all the rows. If you have just a colon that attribute refers to everything. So if you just want to get the first columns using a loc, or iloc syntax. We can write something like df.loc[: , which column we need to refer, to subset the columns.

(Refer Slide Time: 05:49)

- *# subset columns with loc
note the position of the colon
it is used to select all rows*

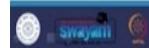
13

The next slide I need to show that we are going to subset the columns with loc, not the position of the colon. It is used to select all rows.

(Refer Slide Time: 05:58)

```
In [45]: subset = df.loc[:, ['year', 'pop']]  
print(subset.head())
```

```
year      pop  
0  1952  8425333.0  
1  1957  9240934.0  
2  1962  10267083.0  
3  1967  11537966.0  
4  1972  13079460.0
```



14

You see that, subset equal df.loc: , I want to see only two columns that is year and population. So when you type this way you will get all the rows only two columns details that is year and population. You will type this so when you type print subset.head. You can get the first 5 rows. So you will see how it appearing.

(Video Starts: 06:26)

Subsets equal to Subset is object because from the df is the initial object which has all the details. Now I am going to fetch only few columns from the df object that I am going to saved in the name subset, subset is the object. So all the rows but I need only year column and population column so I am going to type I want to see the first 5 rows, see that I am able to see 1st 5 rows, only for 2 cells. That is year and population. This is the way to get only 2 cells from the 2 columns from the Big Files.

(Video Ends: 07:21)

(Refer Slide Time: 07:21)

- # subset columns with iloc
- # iloc will allow us to use integers
- # -1 will select the last column

```
In [51]: subset = df.iloc[:, [2, 4, -1]]
print(subset.head())
```

	pop	lifeExp	gdpPercap
0	8425333.0	28.801	779.445314
1	9240934.0	30.332	820.853030
2	10267083.0	31.997	853.100710
3	11537966.0	34.020	836.197138
4	13079460.0	36.088	739.981106

15

There is another example subset column with iloc, iloc will allow us to use integers - 1 will select the last column. The same thing whatever we have seen in the previously so subset equal to df.iloc:, represents all the rows. Then [2 , 4 , -1, then we can see by using this command print subset.head 1st 5 rows.

(Video Starts: 07:46)

See that we are able to see the last column and the population column, life expectancy column. You can open our Excel sheet you can verify whether we are getting the right answer or not.

(Video Ends: 08:26)

(Refer Slide Time: 08:26)

Subsetting Columns by Range

- # create a range of integers from 0 to 4 inclusive

```
In [52]: small_range = list(range(5))
print(small_range)
```

[0, 1, 2, 3, 4]

16

Sometime there is another way for subsetting columns by using the command called range. First will make range of numbers we are going to save that range of a number in object called small _ range, so small _ range equal to list range 5. Print small range will get 0, 1, 2, 3, and 4. Now this small _ range, object can be used to access the corresponding columns.

(Refer Slide Time: 08:57)

- # subset the dataframe with the range

```
In [53]: subset = df.iloc[:, small_range]
print(subset.head())
```

	country	year	pop	continent	lifeExp
0	Afghanistan	1952	8425333.0	Asia	28.801
1	Afghanistan	1957	9240934.0	Asia	30.332
2	Afghanistan	1962	10267083.0	Asia	31.997
3	Afghanistan	1967	11537966.0	Asia	34.020
4	Afghanistan	1972	13079460.0	Asia	36.088

So if I type a subset equal to df.iloc[:, small _ range] I can get.

(Refer Slide Time: 09:04)

Subsetting Columns by Range

- # create a range of integers from 0 to 4 inclusive

```
In [52]: small_range = list(range(5))
print(small_range)
```

[0, 1, 2, 3, 4]

1st column, 2nd column, 3rd column, 4th column and 5th column, so we will try this.

(Video Starts: 09:09)

small_range is an object, we are going to create a range. Suppose we want to see what small_range is. So it is up to 0 to 4, that means 1 to 5. Now we are going to subset using that object called small_range using ilocation command. df.ilocation:small_range we see that here we are able to see 5 column that is a country, year, population, continent and life expectancy.

(Video Ends: 10:21)

(Refer Slide Time: 10:22)

The screenshot shows a Jupyter Notebook cell with the title "Subsetting Rows and Columns". The code in the cell is:

```
In [54]: # using loc
print(df.loc[42, 'country'])
```

The output of the code is "Angola". The notebook interface includes a toolbar at the top and a status bar at the bottom indicating "18".

So far we have seen subsetting only rows and columns. Now we are going to subset rows and columns simultaneously. For example; using loc command so if you type print df.loc 42 countries. We can check in the 42 label in country column. What is the cell name, there cell name is Angola. Will try this.

(Video Starts: 10:47)

Going to see in that file in 42nd label in country column what value is there so that is an Angola, Yes?

(Video Ends: 11:09)

(Refer Slide Time: 11:09)

- # using iloc

```
In [55]: print(df.iloc[42, 0])
```

Angola

19

Yes, we can see what is in the using the same ilocation we can see in 42nd label in 0th column. Now we can represent column also with 0 columns, what value it is, you will see that. You can verify you have to get to the answer. You can open the Excel file. You can verify we are correctly accessing the cell or not.

(Video Starts: 11:29)

Print df.iloc in 42nd label 0th column what is the value it is Angola.

(Video Ends: 11:46)

(Refer Slide Time: 11:46)

Subsetting Multiple Rows and Columns

- #get the 1st, 100th, and 1000th rows
#from the 1st, 4th, and 6th columns

```
In [56]: print(df.iloc[[0, 99, 999], [0, 3, 5]])
```

	country	continent	gdpPerCap
0	Afghanistan	Asia	779.445314
99	Bangladesh	Asia	721.186086
999	Mongolia	Asia	1226.041130

20

Next we can subset multiple rows and columns. For example; get the 1st, 100th and 1000th rows from the 1st, 4th and 6th column. So now we are going to, simultaneously we are going to fetch

rows and columns and corresponding cells. So print to df.iloc 0, 99, 999. Similarly column labels is 0, 3, 5. Let us see what answer.

(Video Starts: 12:13)

This accessing rows and columns are very important functions because nowadays data file comes with a lot of rows and lot of columns. We need not use all the columns, all the rows for further analysis. Sometimes we need only specific rows or specific columns. So these basic commands will help you, how to access a particular rows and columns, that will be very useful when we do further analysis using Python. Yeah? This is the value so that means 1st row, 100th row 1000th row, 1st column and soon.

(Video Starts: 13:08)

(Refer Slide Time: 13:08)

- if we use the column names directly,
it makes the code a bit easier to read
note now we have to use loc, instead of iloc

```
In [57]: print(df.loc[[0, 99, 999], ['country', 'lifeExp', 'gdpPercap']])
```

	country	lifeExp	gdpPercap
0	Afghanistan	28.801	779.445314
99	Bangladesh	43.453	721.186086
999	Mongolia	51.253	1226.041130

And there is another way if you use the column names directly it makes the code a bit easier to read. In terms of number and so you see number column. If you use for representing column, if you use column name we can see what is there, so simply type the column name. So we use this command, print df.loc 0, 99, 999. Then directly will type the column name country, life expectancy, gdpPercap you see there is a square bracket here.

(Video Starts: 13:36)

That you have to do as the same that Life capital Exp, Yes? This is because country, life expectation this is the easy way to because we cannot remember column name.

(Video Ends 14:48)

(Refer Slide Time 14:49)

```
In [58]: print(df.loc[10:13, ['country', 'lifeExp', 'gdpPercap']])
```

	country	lifeExp	gdpPercap
10	Afghanistan	42.129	726.734055
11	Afghanistan	43.828	974.580338
12	Albania	55.230	1601.056136
13	Albania	59.280	1942.284244

22

This was not only that instead of see suppose if you put a 10 column 13 that corresponding rows will be displayed. So print df.loc 10 to 13, the 10th row 11th, row 12th, row 13th, row will be shown and in columns country and life expectancy and gdpPercap so we will try this command.

(Video Starts: 15:11)

That means we can see the range of rows at a time. You are able to see the 10th row, 11th, 12th and 13th.

(Video Ends: 16:17)

(Refer Slide Time: 16:17)

```
In [59]: print(df.head(n=10))
```

	country	year	pop	continent	lifeExp	gdpPercap
0	Afghanistan	1952	8425333.0	Asia	28.801	779.445314
1	Afghanistan	1957	9240934.0	Asia	30.332	820.853030
2	Afghanistan	1962	10267083.0	Asia	31.997	853.100710
3	Afghanistan	1967	11537966.0	Asia	34.020	836.197138
4	Afghanistan	1972	13079460.0	Asia	36.088	739.981106
5	Afghanistan	1977	14880372.0	Asia	38.438	786.113360
6	Afghanistan	1982	12881816.0	Asia	39.854	978.011439
7	Afghanistan	1987	13867957.0	Asia	40.822	852.395945
8	Afghanistan	1992	16317921.0	Asia	41.674	649.341395
9	Afghanistan	1997	22227415.0	Asia	41.763	635.341351

23

Okay? Next see print df. head we can see we can able to see 1st, 10 rows.

(Refer Slide Time: 16:23)

The slide has a dark blue header and footer. The title 'Grouped Means' is centered in a light blue box. Below the title is a bulleted list:

- # For each year in our data, what was the average life expectancy?

To answer this question,
we need to split our data into parts by year;
then we get the 'lifeExp' column and calculate the mean

In the footer, there are three small circular icons and the number 24.

The 10th row some time for each year in our data what was the average life expectancy. To answer this question we need to split our data into parts per year and then we can get the life expectancy column and calculate the mean.

(Refer Slide Time: 16:38)

The screenshot shows a Jupyter Notebook cell with the following content:

```
In [60]: print(df.groupby('year')['lifeExp'].mean())
```

year	lifeExp
1952	49.057620
1957	51.507401
1962	53.609249
1967	55.678290
1972	57.647386
1977	59.570157
1982	61.533197
1987	63.212613
1992	64.160338
1997	65.014676
2002	65.694923
2007	67.007423

Name: lifeExp, dtype: float64

In the footer, there are three small circular icons and the number 25.

So what is happening there is a command which I go to use called groupby, we look at the data it is not grouped. So when you use this command print df.groupby year, and life expectancy and corresponding mean. The mean of the on the in the year 1952, the mean of the life expectancy variable is 49.05. In 57, 51.09. We look at the data; it is not in this order. So the groupby by year

this command is grouping all the values, with respect to year. So we will see what is the answer for this, we will verify this.

(Video Starts: 17:15)

When you open that Excel file you will see that the Excel file will be in some other form it is not grouped by year, different years are appearing at different places. So this command that is a group by will help you to group the data in year wise. Yes, you see that you are able to get 1952 the life expectancy was 49 years you see that when you look at this data. When year increases the life expectancy year also increases due to advancement of medical facility available and the standard of life is also increasing.

(Video Ends: 18:42)

(Refer Slide Time: 18:42)

```
In [61]: multi_group_var = df.\n        groupby(['year', 'continent'])\n        [['lifeExp', 'gdpPerCap']].\n        mean()\n        print(multi_group_var)
```

year	continent	lifeExp	gdpPerCap
1952	Africa	39.135500	1252.572466
	Americas	53.279840	4079.062552
	Asia	46.314394	5195.484004
	Europe	64.408500	5661.057435
	Oceania	69.255000	10298.085650
1957	Africa	41.266346	1385.236062
	Americas	55.960280	4616.043733
	Asia	49.318544	5787.732940
	Europe	66.703067	6963.012816
	Oceania	70.295000	11598.522455
1962	Africa	43.319442	1598.078825
	Americas	58.398760	4981.541870
	Asia	61.662772	6770.260676

Now, we can form a stacked table. Stacked table is using the group by command. So you type this multi_group_variable = df . \ . See the \ represents to breaking the command we can use \ . Otherwise you can write straightaway also no problem. df.groupby(year, continent, life_expectancy,gdp_per_capita), then we can find the mean. Then we will get this output for that means in 1952, in Africa, the life expectancies 39, in America 53, in Asia 46 in Europe 64 will try this command.

(Video Starts: 19:28)

When we takes these command you will get an output, that is a stacked table. That is very useful for interpreting the whole dataset, is kind of a way of summarizing the data in the form of table.

Multi_group. You see that now year wise. It is very, very useful command it is year by 1952, some country Africa. What was the average year 1957 Africa. We see that if you look at only the Africa data. 52 to 39 in 57 41, in 62 43, in 67 45, see that we can interpret this way, by looking at the, this table. Suppose you have to flatten this.

(Video Ends 21:24)

(Refer Slide Time: 21:24)

- If you need to “flatten” the dataframe, you can use the `reset_index` method.

```
In [42]: flat = multi_group_var.reset_index()
print(flat.head(15))

   year continent  lifeExp  gdpPerCap
0  1952      Africa  39.135500  1252.572466
1  1952      America  53.279840  4079.062552
2  1952        Asia  46.314394  5195.484004
3  1952      Europa  64.408500  5661.057435
4  1952    Oceania  69.255000  18298.855500
5  1957      Africa  41.266346  1385.236062
6  1957    Americas  55.960280  4616.043733
7  1957        Asia  49.318544  5787.732940
8  1957      Europa  66.703067  6963.012816
9  1957    Oceania  70.295000  11598.522455
10 1962      Africa  43.319442  1598.078825
11 1962    Americas  58.398760  4901.541870
12 1962        Asia  51.563223  5729.369425
13 1962      Europa  68.539233  8365.406614
14 1962    Oceania  71.085000  12696.452430
```

27

If, you need to flatten the data frame. You can use this reset underscore index method, just to type `flat = multi_group_var.reset_index`. Then you see now the data is again. Now it is flattened. The same data set, which was it in the table form now it in the simple learned form. So we will try this comment.

(Video Starts 21:48)

This is what you are doing the data manipulation, because from the big data file, we have to learn this kind of fundamental data manipulation methods that will be very useful, in coming classes. So able to use `reset_index` command to flatten the, that stacked table. See that now we can see first 15 rows. Now it is data is flattened into the normal form.

(Video Ends 22:41)

(Refer Slide Time: 22:41)

Grouped Frequency Counts

- use the nunique to get counts of unique values on a Pandas Series.

```
In [63]: print(df.groupby('continent')['country'].nunique())
```

continent	country
Africa	52
Americas	25
Asia	33
Europe	30
Oceania	2

Name: country, dtype: int64

28

The next one is grouped frequency counts. By using nunique command, we can get a count of unique values on the panda series. So when you type print df. groupby continent, country. nunique, you can get unique values that means frequency. Okay, will try this command.

(Video Starts: 23:04)

Print, See Africa 52, America is 25, Asia 33. When you look at the data, again, you go to excel, Excel data you can interpret what is the 52 means, what is the America 25 and soon.

(Video Ends: 23:49)

(Refer Slide Time: 23:49)

Basic Plot

```
In [65]: global_yearly_life_expectancy = df.groupby('year')['lifeExp'].mean()
print(global_yearly_life_expectancy)
```

year	lifeExp
1952	49.057620
1957	51.507481
1962	53.609249
1967	55.678298
1972	57.647386
1977	59.570157
1982	61.533197
1987	63.212613
1992	64.160338
1997	65.014676
2002	65.694923
2007	67.007423

Name: lifeExp, dtype: float64

29

Now, some basic plot a way to construct two things one is year and life expectancy. So we are going to create a new object that is called Global _ yearly_ life _expectancy. By grouping year

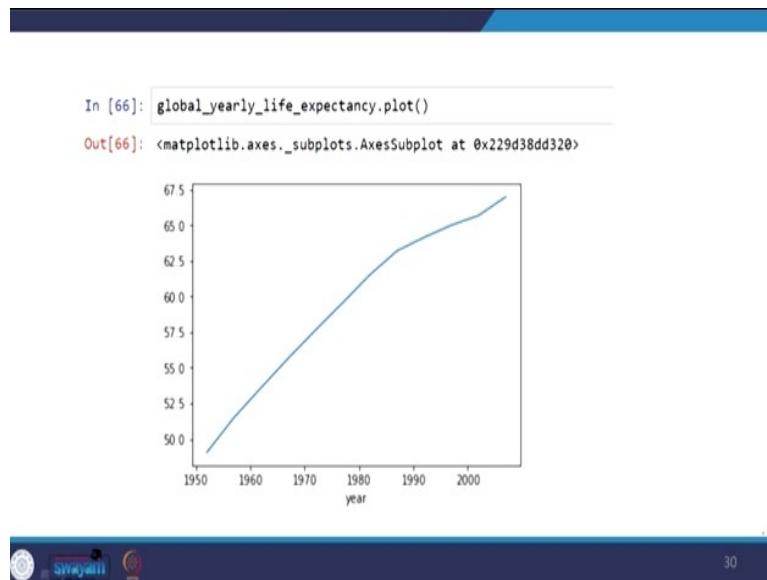
and life expectancy, with respect to its mean. Then we are going to print it. So you are going to get two values one is year. Next one is life expectancy. That is a mean life expectancy, you will see this.

(Video Starts: 24:17)

There is a new object. The object name is called Global _ yearly _ life expectancy. Yes, see that year, and supposed we want to plot it. We will see we are going to plot this data, how we are going to plot it.

(Video Ends: 25:28)

(Refer Slide Time: 25:28)



Simply, just that object name. plot. That automatically takes this was output, which I got is in x axis in a year, in y axis, average life expectancy. We will run this.

(Video Starts 25:40)

So, what this data says that, when the year 1950 - 2000 you see when the year increases, the life expectancy also increases.

(Video Ends: 26:07)

(Refer Slide Time 26:07)

Visual Representation of the Data

- Histogram -- vertical bar chart of frequencies
- Frequency Polygon -- line graph of frequencies
- Ogive -- line graph of cumulative frequencies
- Pie Chart -- proportional representation for categories of a whole
- Stem and Leaf Plot
- Pareto Chart
- Scatter Plot

31

Just we have seen only the simple plot, in coming classes, we will see some of the visual representation of the data. We are going to see a histogram, frequency polygon, ogive curves, pie chart, stem and leaf plot and pareto chart and scatter plot .

(Refer Slide Time: 26:21)

Methods of visual presentation of data

- Table

	1st Qtr	2nd Qtr	3rd Qtr	4th Qtr
East	20.4	27.4	90	20.4
West	30.6	38.6	34.6	31.6
North	45.9	46.9	45	43.9

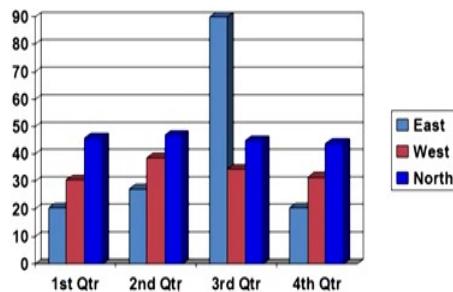
32

Suppose, this is the data, see what is there in East, west, north. In column first quarter, second quarter, third quarter, fourth quarter.

(Refer Slide Time: 26:30)

Methods of visual presentation of data

- Graphs



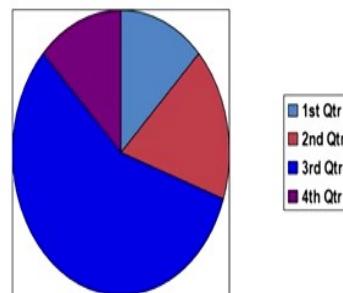
33

Suppose the very easiest way is the graph. By using this is called bar graph, bar chart. Bar chart is different regions are labeled as different colors. This is a method of visual representation of the data. If you look at this, the eastern side in third quarter, there are more sales. Okay.

(Refer Slide Time: 26:53)

Methods of visual presentation of data

- Pie chart



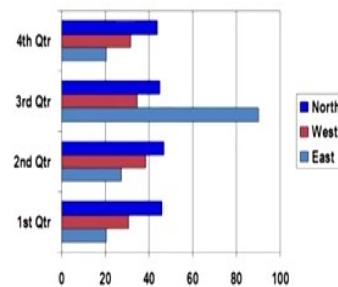
34

The another way to represent visually, the data is pie chart, is the first quarter, third quarter. You look at this, third quarter, which is in blue in color. There are more sales. And most importantly the pie chart, we can get pie chart only for categorical variable. The variable is continuous, you cannot use bar chart, you cannot use pie chart. So the pie chart is used only for categorical variable. That is for only count data.

(Refer Slide Time: 27:31)

Methods of visual presentation of data

- Multiple bar chart



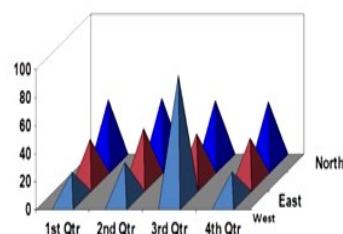
35

The another one is the Multiple bar chart. This is another way to represent the data visually.

(Refer Slide Time: 27:39)

Methods of visual presentation of data

- Simple pictogram



36

Another one is a simple pictogram.

(Refer Slide Time: 27:43)

Frequency distributions

- Frequency tables

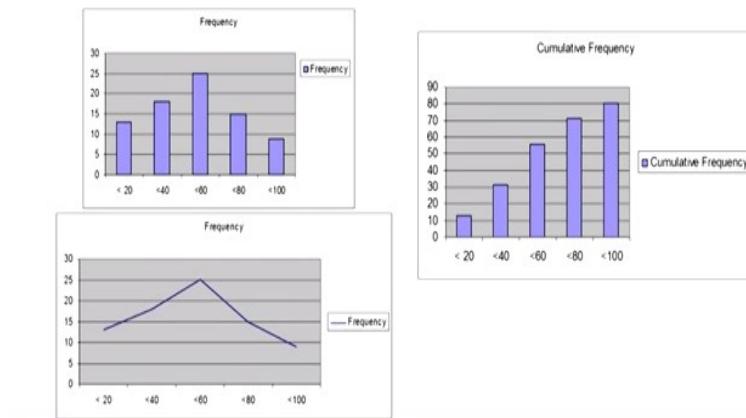
Observation Table		
Class Interval	Frequency	Cumulative Frequency
< 20	13	13
< 40	18	31
< 60	25	56
< 80	15	71
< 100	9	80

37

See, this is the frequency table.

(Refer Slide Time: 27:25)

Frequency diagrams



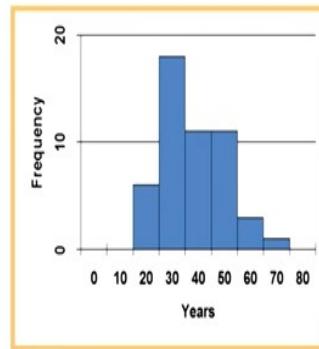
38

See, next one is frequency polygon. This figure is drawn from the previous table, which was shown in the previous slide. So below 20 around 13,14. This represents frequency polygon. When you connect the midpoint, you see that this is the. This is called frequency polygon. Then the, this one is the cumulative frequency. It is not always, you cannot connect the midpoint, you have to be very careful with the data is continuous, then only you can connect one this bar. The data is not continuous, you cannot connect it.

(Refer Slide Time: 28:24)

Histogram

Class Interval	Frequency
20-under 30	6
30-under 40	18
40-under 50	11
50-under 60	11
60-under 70	3
70-under 80	1



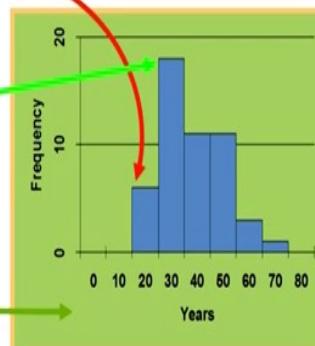
39

Next one is a histogram .The histogram was constructed from the given table. You see.

(Refer Slide Time: 28:30)

Histogram Construction

Class Interval	Frequency
20-under 30	6
30-under 40	18
40-under 50	11
50-under 60	11
60-under 70	3
70-under 80	1



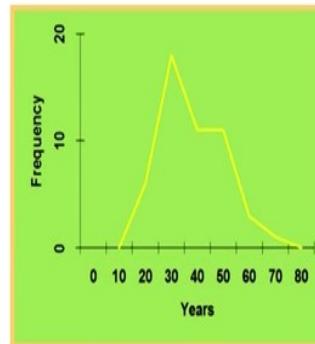
40

The lower limit of the table values is going to in x axis. The frequency is shown in the y axis. You see that this is data in continuous data. Okay, that was histogram. The purpose of histogram is, the histogram will give you a rough idea what is the nature of the data whether, what kind of distribution it follows. Whether it is following bell shaped curve, whether the data is skewed right or skewed left.

(Refer Slide Time: 29:03)

Frequency Polygon

<u>Class Interval</u>	<u>Frequency</u>
20-under 30	6
30-under 40	18
40-under 50	11
50-under 60	11
60-under 70	3
70-under 80	1



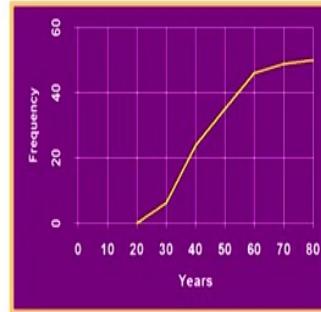
41

Next one is the frequency polygon which I have shown you. If, the midpoint of histogram are connected then there is called frequency polygon. Because, the frequency polygon is used to know the trend.

(Refer Slide Time: 29:20)

Ogive

<u>Class Interval</u>	<u>Cumulative Frequency</u>
20-under 30	6
30-under 40	24
40-under 50	35
50-under 60	46
60-under 70	49
70-under 80	50



42

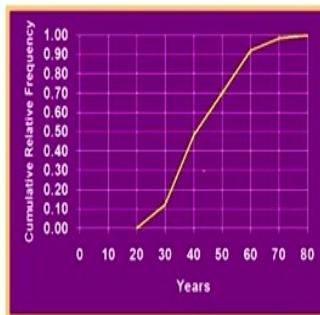
Trend of the data. The next one is ogive curve. This is cumulative frequency curve .So what is happening in the, for example 20- under 30, the upper limit of the interval is taken the x axis, the cumulative frequency is taken in the y axis. For example, the first interval.20 - 30. So 30 the upper interval is 6. For 40, upper interval is to 24, that is marked.

Because the advantage of this ogive curve is, supposed if we want to know below 16, how many numbers are there, that can be read directly from the ogive curve. That is the purpose of ogive curve.

(Refer Slide Time: 29:56)

Relative Frequency Ogive

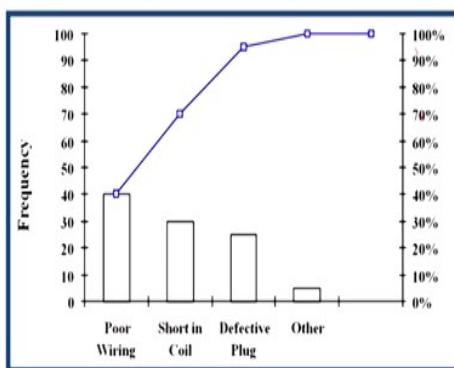
<u>Class Interval</u>	<u>Cumulative Relative Frequency</u>
20-under 30	.12
30-under 40	.48
40-under 50	.70
50-under 60	.92
60-under 70	.98
70-under 80	1.00



Next one is the relative frequency curve. Exactly similar to that now actual frequency that relative frequency was taken.

(Refer Slide Time: 30:08)

Pareto Chart



Okay. The next way to represent the data using pareto chart. The Pareto chart is having some applications in quality control also. This is to identify which is more important, important variable. Assume that, if you look at this Pareto chart. There are 3 axes one is frequency. In x

axis, different name is given poor wiring, short in coil, defective plug, other. You see there is one more variable in terms of percentage.

For example, I am a quality control engineer, suppose my motor is failing so often. I want to know there are different reason for failing of the motor. I want to know what are the main reasons, due to which the motor fails. So what I have done. First I have go to frequency table, that is due to poor wiring, the motor was failing for failing 40 times, frequencies 40. Due to short in coil, the motor was failed 30 times.

Due to defective in plug, the motor was failed 25 times. Due to some other reasons the motor was failed by say below 10 times. So the first technique is for drawing this one, we have to arrange in the descending order of their frequency. So in x axis that values are taken. Then the cumulative frequencies plotted on the, this axis. For example, how to interpret this table is. You see, here this value corresponding this only 70.

So 70 % of the failure is due to only two reasons, that is poor wiring and short circuiting. So what is the meaning of this one is, if you are able to address these 2 problems, 70% of the failures can be eliminated. So the purpose of a Pareto chart is, to identify which is critical for us. Generally it is called 80-20 principle. This is called the Pareto principle .That is 80% of the problems are due to 20% of the reasons.

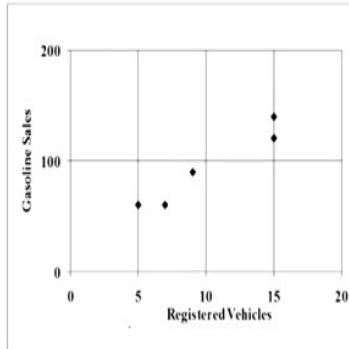
So similarly here, when you look at this, the cell, here need not always 80, see the 70% the failures, only due to 2 factors that is due to poor wiring and short coil. So this is the pareto chart.

(Refer Slide Time: 32:33)

Scatter Plot

Registered Vehicles (1000's)	Gasoline Sales (1000's of Gallons)
---------------------------------	---------------------------------------

5	60
15	120
9	90
15	140
7	60



45

The next one is scatter plot. The scatter plot is so far what ever seen only for one variable, the scatter plot is used for two variable. In x axis registered vehicle, y axis the gasoline sales. So this says the scatter plot says, when the number of registered vehicle is increasing the gasoline sales is also increasing. So the scatter plot is used to know the trend out the data.

(Refer Slide Time: 32:59)

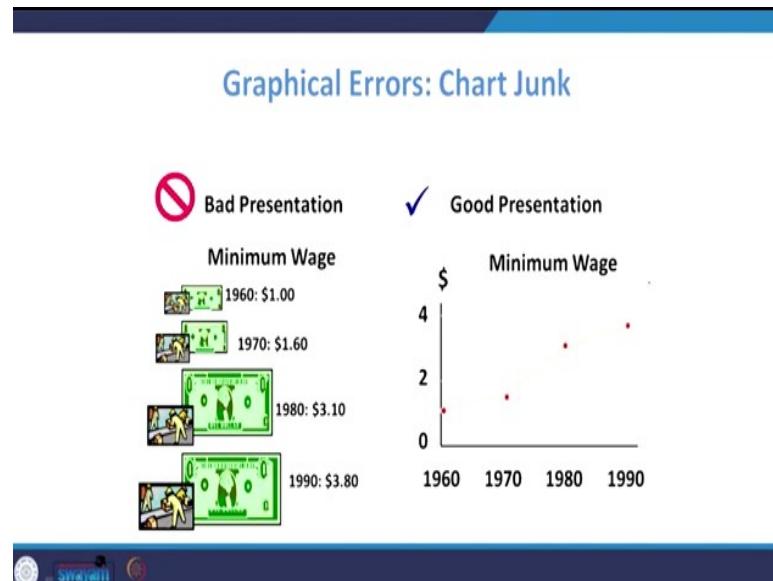
Principles of Excellent Graphs

- The graph should not distort the data
- The graph should not contain unnecessary adornments (sometimes referred to as chart junk)
- The scale on the vertical axis should begin at zero
- All axes should be properly labeled
- The graph should contain a title
- The simplest possible graph should be used for a given set of data

Some of the basic principle for excellent graph. One is the graph should not distort the data. The graph should be very simple. It should not contain unnecessary adornments. So, so much decoration in the graph is not required, the scale on the vertical axis should begin at 0. All axes should be properly labeled. Whether should be x axis or y axis, it has to be properly labeled. The

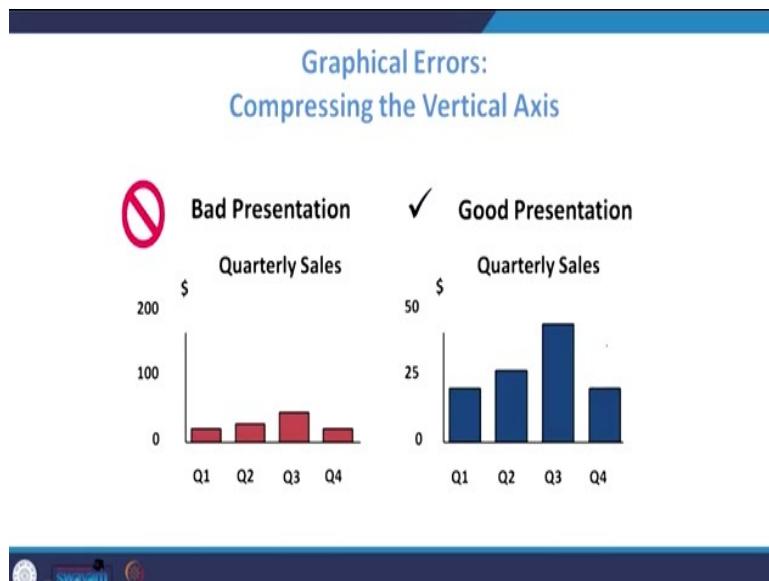
graph should contain a title. The simplest possible graph should be used for given set of data. These are the basic principle of excellent graph.

(Refer Slide Time: 33:39)



See when you look at this one. The left hand side it is a bad representation of the graph. What is happening lot of animations, unnecessary pictures. The right hand side, it is a simple graph x axis is taken as year, in y axis it has taken the wage. So it is showing some trend. But when you look on the left hand side it is not giving any idea. What is happening year with respect to wage.

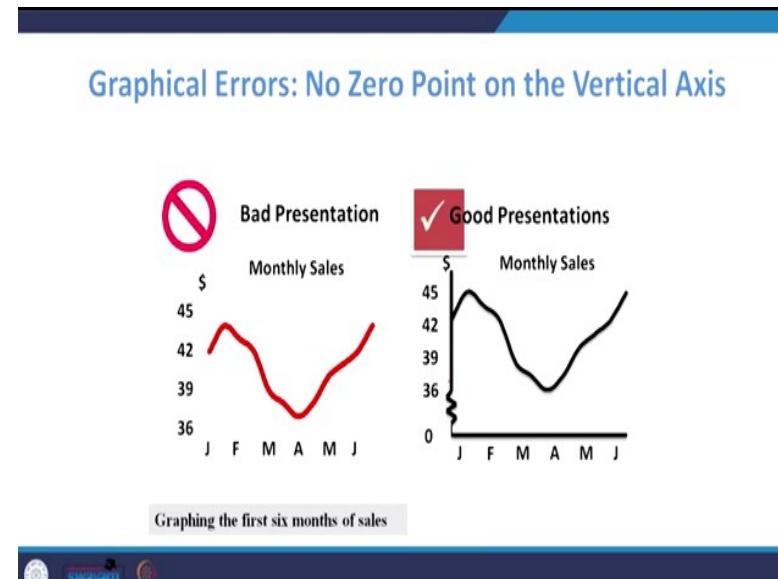
(Refer Slide Time: 34:04)



Another one you look at the left side picture and right side picture. Both are the same data. But what is happening. When here in the left side picture the scale is 0 to 100, here it is 0 to 25 just

by changing the scale, we are able to get different interpretation. You see that when the scale is increased. It looked like flat. If you are drawing in smaller scale. You see that look like there's a lot of variations. So what is the learning is that we are to use proper scale to draw the picture.

(Refer Slide Time: 34:40)



The next one is the graphical error, no 0 point on the vertical axis. When you look at the left side of the figure January, February, March, April, May, June, the month is given in x axis. Monthly sales is given y axis. But the problem on the left hand side is it did not start from 0. The right side is you see that the small Brake is given. So, that, even though, 0 to 36 there is no data, you have to make a small break like this. So that, we can come to know it start from 0.

So this is the right hand side is the right way of drawing the graph. This is the basic requirement. In this lecture, what you have seen, how to access particular rows and columns by using basic commands. Then we have seen the different visualization techniques, different theories of the visualization technique. The next class will take in some sample data. By using the sample data with the help of the sample data will try to visualize the data.

By having different tools like a pie chart, bar chart, pictogram, Pareto chart or simple graph. Thank you, we will see you next class.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology, Roorkee

Lecture No 4
Central Tendency and Dispersion

Good morning students, today we are going to the lecture 4. In this lecture we are going to talk about central tendency and how to measure the dispersions. The lecture objectives we talk about different types of central tendency.

(Refer Slide Time: 00:42)

Lecture objectives

- Central tendency
- Measures of Dispersion

Then different types of dispersions

(Refer Slide Time: 00:44)

Measures of Central Tendency

- Measures of central tendency yield information about “particular places or locations in a group of numbers.”
- A single number to describe the characteristics of a set of data

What is measure of central tendency? Measure of central tendency yield information about particular places or locations in a group of numbers. Suppose there are a group of number is there that number group of numbers has to be replaced by a single number that single number we can call it as central tendency. That is a single number to describe the characteristics of a set of data.

(Refer Slide Time: 01:08)

Summary statistics

- Central tendency or measures of location
 - Arithmetic mean
 - Weighted mean
 - Median
 - Percentile
- Dispersion
 - Skewness
 - Kurtosis
 - Range
 - Interquartile range
 - Variance
 - Standard score
 - Coefficient of variation



4

Some of the central tendency which we are going to see in this lecture is arithmetic mean, weighted mean, median, and percentile. In the dispersions we are going to talk about skewness, kurtosis, range, interquartile range, variance, standard score and coefficient of variation.

(Refer Slide Time: 01:25)

Arithmetic Mean

- Commonly called 'the mean'
- It is the average of a group of numbers
- Applicable for interval and ratio data
- Not applicable for nominal or ordinal data
- Affected by each value in the data set, including extreme values
- Computed by summing all values in the data set and dividing the sum by the number of values in the data set



5

First we will see the first central tendency arithmetic mean. Commonly it is called as the mean it is the average of a group of numbers; it is applicable for interval and ratio data. This point is very important it is not applicable for nominal and ordinal data. It is affected by each value in the data set including extreme values, one of the problem of the with the mean is that it is affected by the extreme values computed by summing all values in the data set and dividing the sum by the number of values in the data set.

(Refer Slide Time: 02:01)

Population Mean

$$\begin{aligned}\mu &= \frac{\sum X}{N} = \frac{X_1 + X_2 + X_3 + \dots + X_N}{N} \\ &= \frac{24 + 13 + 19 + 26 + 11}{5} \\ &= \frac{93}{5} \\ &= 18.6\end{aligned}$$

See here I have used a notation μ , μ means capital letters μ represents, mean for the

population. The formula $\mu = \frac{\sum X}{N}$
 $= (X_1 + X_2 + X_3 + \dots + X_N) / N$

here N is the number of elements. For example; the values are 24, 13, 19, 26, 11 add these numbers and divided by 5 because there are 5 elements. So $93 / 5$, 18.6 is the mean of these 5 numbers. So now the 18.6 can be replaced by these set 5 numbers. Okay?

Suppose in your class if you see the average mark is 60. So the whole marks of all the students can be represented to be a single number that is 60, 60 will give an idea about the performance of the whole class.

(Refer Slide Time: 02:55)

Sample Mean

$$\begin{aligned}\bar{X} &= \frac{\sum X}{n} = \frac{X_1 + X_2 + X_3 + \dots + X_n}{n} \\ &= \frac{57 + 86 + 42 + 38 + 90 + 66}{6} \\ &= \frac{379}{6} \\ &= 63.167\end{aligned}$$



7

Next what is the sample mean? Make sure that the difference here the \bar{X} . Previously for the population mean we have used μ for the sample mean we are using \bar{X} .

$$\bar{X} = \frac{\sum X}{N}$$

$$= X_1 + X_2 + X_3 / n.$$

For example; 6 element is there 57, 86, 42, 38, 19, 66 so divided by 6, the mean is 63.167

(Refer Slide Time: 03:19)

Mean of Grouped Data

- Weighted average of class midpoints
- Class frequencies are the weights

$$\begin{aligned}\mu &= \frac{\sum fM}{\sum f} \\ &= \frac{\sum fM}{N} \\ &= \frac{f_1M_1 + f_2M_2 + f_3M_3 + \dots + f_nM_n}{f_1 + f_2 + f_3 + \dots + f_n}\end{aligned}$$



8

Now how to find out the mean of a grouped data? The mean of your grouped data is nothing but weighted average of class midpoints, class frequencies are the weight. For the formula is

$$\mu = \frac{\sum fM}{\sum f}$$

So Sigma f is nothing but

= (f₁ m₁ + f₂ m₂ + f₃ m₃ and so on + f_i m_i)/ sum of all f.

That is nothing but your N. We will see you an example;

(Refer Slide Time: 03:58)

Calculation of Grouped Mean

Class Interval	Frequency(f)	Class Midpoint(M)	fM
20-under 30	6	25	150
30-under 40	18	35	630
40-under 50	11	45	495
50-under 60	11	55	605
60-under 70	3	65	195
70-under 80	1	75	75
	50		2150

$$\mu = \frac{\sum M}{\sum f} = \frac{2150}{50} = 43.0$$



See this is the grouped data. What is given class interval is given frequency is given class midpoint is given and multiplied value of frequency and midpoint also we can find out. For example; see here 20 to 30 there are 6 numbers is their frequency 6. Suppose if you say the marks of here if you say this this is an example of here marks obtained by in your class. So between 20 and 30 there are 6 students is there. Between 30 under 40 there are 18 students is there.

Suppose for this the data is in this format that is in grouped format how to find out the mean? Okay? First what do you have to do first you have to find out the class midpoint. See 20 to 30 that is a class interval the midpoint is 25, for 30 and 40. The class midpoint is the middle value 35 like this 45, 55, 65, and 75. Next one you have two multiplied by frequency and class midpoint so 6 into 25 is 150, 18 into 35 is 630, 11 into 45 is 495 and so on.

$$\frac{\sum fM}{\sum f}$$

What the formula says it is $\frac{\sum fM}{\sum f}$ last column the sum value is 2150, 2150/50

Sigma f is some of the frequency so for this kind of grouped data the mean is 43.

(Refer Slide Time: 05:29)

Weighted Average

- Sometimes we wish to average numbers, but we want to assign more importance, or weight, to some of the numbers.
- The average you need is the weighted average.



Now we will go to the next central tendency that is the weighted average. Some time if you look at the previous values, the each value is given equal weightage. Suppose it is not always the case there may be some marks there some values where there may be higher weightage. So for that case we have to go for weighted average. Some time you see this we will list two average numbers but we want to assign more importance or weight to some of the numbers. The average you need is the weighted average.

(Refer Slide Time: 05:58)

Formula for Weighted Average

$$\text{Weighted Average} = \frac{\sum xw}{\sum w}$$

where x is a data value and w is the weight assigned to that data value. The sum is taken over all data values.



So the weighted average is sum of the product of weightage and that value/sum of values. Where x is the data value and w is the weight assigned to that data value. The sum is taken over all data values.

(Refer Slide Time: 06:20)

Example

Suppose your midterm test score is 83 and your final exam score is 95. Using weights of 40% for the midterm and 60% for the final exam, compute the weighted average of your scores. If the minimum average for an A is 90, will you earn an A?

$$\begin{aligned} \text{Weighted Average} &= \frac{(83)(0.40)+(95)(0.60)}{0.40+0.60} \\ &= \frac{32+57}{1} = 90.2 \end{aligned}$$

You will earn an A!

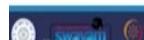


We will see one application of this weighted average. Suppose your midterm test score is 83 and your final exam is score is 95 using weights of 40% is for the midterm and 60% is for the final exam compute the weighted average of your scores if the minimum average for an A grade is 90 will you earn an A grade. So first we find out the weighted average so the mark is 83 weights age is 40% for midterm for interim your mark is 95 weightage is 60 %.

So multiply that then divided by some of the weight that = $0.4 + 0.6 = 1$, so 90.2. So if you are above 90 you will get A, because you are crossing 90 obviously you will get the A grade.
(Refer Slide Time: 07:12)

Median

- Middle value in an ordered array of numbers
- Applicable for ordinal, interval, and ratio data
- Not applicable for nominal data
- Unaffected by extremely large and extremely small values



13

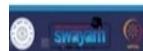
Now we will go to the next central tendency Median, the middle value in ordered array of number is called Median. It is applicable for ordinal interval and ratio data. You see previously the mean is applicable only for interval and ratio data but the median is applicable

for ordinal data. There is a point has to be remembered and it is not applicable for nominal data and one advantage of median is it is unaffected by extremely large and extremely small values.

(Refer Slide Time: 07:46)

Median: Computational Procedure

- First Procedure
 - Arrange the observations in an ordered array
 - If there is an odd number of terms, the median is the middle term of the ordered array
 - If there is an even number of terms, the median is the average of the middle two terms
- Second Procedure
 - The median's position in an ordered array is given by $(n+1)/2$.



14

Next we will see how to compute median there are 2 procedures, first procedure is arrange the observations in an ordered array. If there is an odd number of a term the median is the middle term of the ordered array. If there is the even number of terms the median is the average of middle two terms. Another procedure is the medians position of an ordered array is given by $n + 1 / 2$, n is the number of data set.

(Refer Slide Time: 08:15)

Median: Example with an Odd Number of Terms

Ordered Array

3 4 5 7 8 9 11 14 15 16 16 16 17 19 19 20 21 22

- There are 17 terms in the ordered array.
- Position of median = $(n+1)/2 = (17+1)/2 = 9$
- The median is the 9th term, 15.
- If the 22 is replaced by 100, the median is 15.
- If the 3 is replaced by -103, the median is 15.



15

We will see this example; I have taken some exam some numbers that is arranged in an order that is an ascending order 3, 4, 5, 7 up to 22. There are 17 terms in the ordered array the

position of the median is, with respect to previous let $n + 1 / 2$. So $n + 1 / 2 = (17+1)/2 = 18 / 2 = 9$. So the median is the 9th term, 9th term here is 15. If see the 22 which is the highest number is replaced by 100 still the median is 15.

See if the 3 is replaced by -103 still the median is 15. So there is the advantage of this median over mean is median is not disturbed by extreme values.

(Refer Slide Time: 08:59)

Median: Example with an Even Number of Terms

Ordered Array

3 4 5 7 8 9 11 14 15 16 16 16 17 19 19 20 21

- There are 16 terms in the ordered array
- Position of median = $(n+1)/2 = (16+1)/2 = 8.5$
- The median is between the 8th and 9th terms, 14.5
- If the 21 is replaced by 100, the median is 14.5
- If the 3 is replaced by -88, the median is 14.5



16

Previously the number of items are odd now let us see the another situation; there are 16 terms in the ordered array there is an even number, the position of the median is $n + 1 / 2$ that is $16 + 1 / 2$ is 8.5. So we have to look at the term where it is the position of 8.5. That is the median is between 8th and the 9th term here the 8th term is 14, 9th term is 15 so average of that one is 14.5. Again, if the 21 is replaced by 100 the median is same 14.5, if the 3 is replaced by -88 still the median is 14.5.

(Refer Slide Time: 09:42)

Median of Grouped Data

$$Median = L + \frac{\frac{N}{2} - cf_p}{f_{med}} (W)$$

Where :

L = the lower limit of the median class

cf_p = cumulative frequency of class preceding the median class

f_{med} = frequency of the median class

W = width of the median class

N = total of frequencies



17

Now let us see how to find out the median of your grouped data but it will be grouped data, here if the data is given in the form of a frequency table. This case the formula to find out the

median of a group data is median = $L + (W) \frac{\frac{N}{2} - cf_p}{f_{med}}$. Where, L is the lower limit of the median class before using this formula from the given table you have to find out what is the median class.

Then cf_p = cumulative frequency of the class preceding the median class f median, the frequency of the median class, W is the width of the median class; N is the total number of frequencies.

(Refer Slide Time: 10:26)

Median of Grouped Data -- Example

Class Interval	Frequency	Cumulative Frequency	
20-under 30	6	6	
30-under 40	18	24	
40-under 50	11	35	
50-under 60	11	46	
60-under 70	3	49	
70-under 80	1	50	
$N = 50$			
			$Md = L + \frac{\frac{N}{2} - cf_p}{f_{med}} (W)$
			$= 40 + \frac{50 - 24}{11} (10)$
			$= 40.909$

See this is an example; as I told you before using this formula first order to find out the median class. What is the median class is when you add the frequency it is a 50. $6 + 18 + 11 + 11 + 3 + 1$ is 50. So divide this $50 / 2$ it is a 25. In the community frequency column in the last column look at where that 25 is lying it is not between 30 and 40; it is going to lie on between 40 and 50 because 24 for the next term is 35.

So the median class for this given group data is 40 and 50. So as usual L, is the lower limit of the median class that is a $40 + N$ is 50. You see the cumulative frequency of the preceding interval is 24. So,

$$Md = 40 + ((50/2) - 24) \times 10 / 11$$

because the width interval is 10. When you simplify you would get 40.909, so this is the way to find out the median of your grouped data.

(Refer Slide Time: 11:45)

Mode

- The most frequently occurring value in a data set
- Applicable to all levels of data measurement (nominal, ordinal, interval, and ratio)
- Bimodal -- Data sets that have two modes
- Multimodal -- Data sets that contain more than two modes



19

Now mode the most frequently occurring value in a data set is mode applicable to all level of data, measurement nominal, ordinal, interval and ratio. Sometimes there is a possibility the data set may be bimodal. Bimodal means data sets that have two modes. That means two numbers are repeated same number of time multimodal data sets that contain more than two modes.

(Refer Slide Time: 12:12)

Mode -- Example

- The mode is 44
- There are more 44s than any other value

35	41	44	45
37	41	44	46
37	43	44	46
39	43	44	46
40	43	44	46
40	43	45	48



20

See this one sample data as it is given for this data set the mode is 44, because the 44 is appearing more number of time. How many number of time 1, 2, 3, 4, 5. Okay? So the mode is 44. That is there are more 44s than any other values.

(Refer Slide Time: 12:37)

Mode of Grouped Data

- Midpoint of the modal class
- Modal class has the greatest frequency

Class Interval	Frequency
20-under 30	6
30-under 40	18
40-under 50	11
50-under 60	11
60-under 70	3
70-under 80	1

$$\text{Mode} = L_{Mo} + \left(\frac{d_1}{d_1 + d_2} \right) w = \\ 30 + \left(\frac{12}{12+7} \right) 10 = 36.31$$



21

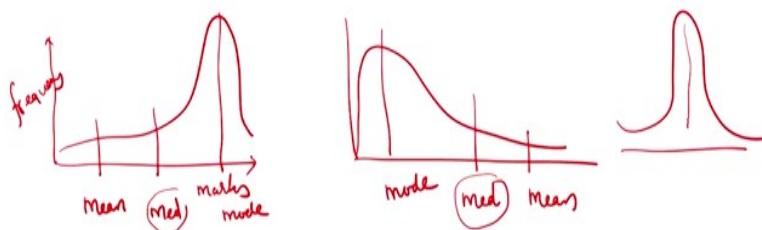
That is the formula for finding mode of a grouped data. Here first we have to find out the mode class. For that look at the frequency column there 18 is the highest frequency. So corresponding the n class interval is called mode interval. Okay? The mode interval L Mo is the lower limit of that mode interval is $= 30 + (12/(12+7)) \times 10$

And d2 is difference between 18 and 11.

See 30 + see d1 is nothing but 18 is the mode interval and then the previous frequency is 6, so $18 - 6$ is 12 / d1 is 12 + d2 is the difference between your 18 and 11 that is 7. So $12 + 7$ multiplied by width is 10, so 36.31 is the mode of your grouped data. Yes? We have studied mean, median, mode for group data and ungrouped data. Now the question is when to use mean? When to use median? When to use mode? Okay?

Many time even though we study mean, median, mode we are not exactly told how to use or when to use mean or when to use median or when to use mode?

(Refer Slide Time: 14:00)



For example look at this data set, this is left skewed data because the tail is on the left hand side. The example for this is suppose, say the exam is very easy question paper and the x axis is the marks and y axis is frequency. So there is more number of students who got higher marks .Where the question paper is easy situation this is an example of left skewed data. So what will happen here, here will be mean here will be median this will be mode.

You see another example; where the question paper is very tough. So this is called right skewed data. You know here what is happening how we are saying that since the question paper is very tough. There are more number of students who got the lesser marks that is why the skewness on this side. So here there will be mean here will be median this will be mode Okay? There will be another situation it is symmetric it is a bell looking at bill shaped curve in this situation.

Now after looking at this hypothetical problem now the question arises when to use mean, when to use median, mode look at the location of the median. The median is always in the middle. Whether the data is left skewed or right skewed the median is always the middle. So whenever the data is skewed you should go for median as a central tendency. If your data is following a bell-shaped curve then you can use mean, median, mode.

There is no problem at all the clue for that choosing the correct central is first you have to plot that curve go to plot the data outer plotting the data you have to get an idea of the skewness of the data set. How it is distributed? Whether it is right skewed or left skewed or it is bell shaped curve. If it is skewed data you go for median as the center tendency. If it is following a bell-shaped curve you go for mean or median or mode as a central tendency.

(Refer Slide Time: 16:39)

Percentiles

- Measures of central tendency that divide a group of data into 100 parts
- Example: 90th percentile indicates that at most 90% of the data lie below it, and at least 10% of the data lie above it
- The median and the 50th percentile have the same value
- Applicable for ordinal, interval, and ratio data
- Not applicable for nominal data



23

Now you go to next one is a Percentile, mainly this you might have seen some of the cat examination scores or gate examination scores their performance is expressed in terms of percentile not the percentage because percentile is having some advantage over percentage because percentage is absolute term but the percentile is the relative term the measure of central tendency that divide a group of data into 100 parts it is called percentile.

For example; somebody say 90th percentile my score is 90th percentile indicates that at most 90% of the data lie below it and at least 10% the data lie above it. Okay? The median and the 50th percentile have the same value. It is applicable for ordinal, interval and ratio data it is not applicable for nominal data.

(Refer Slide Time: 17:44)

Percentiles: Computational Procedure

- Organize the data into an ascending ordered array
- Calculate the p th percentile location:
$$i = \frac{P}{100}(n)$$
- Determine the percentile's location and its value.
- If i is a **whole number**, the percentile is the average of the values at the i and $(i+1)$ positions
- If i is **not a whole number**, the percentile is at the $(i+1)$ position in the ordered array



24

Okay we will see an example how to compute a percentile the first step is organize the data into an ascending ordered array calculate the p th percentile location. Suppose if I want to know 30th percent location for that you have to find out the value i , $i = (P / 100)$ multiplied by n , n is the number of data set, the i is nothing but the percentiles location we got to find out the i value.

If i is a whole number the percentage is the average of the values at the i and $i + 1$ positions.

If i is not a whole number the percentile is the $i + 1$ position in the ordered array.

(Refer Slide Time: 18:35)

Percentiles: Example

- Raw Data: 14, 12, 19, 23, 5, 13, 28, 17
- Ordered Array: 5, 12, 13, 14, 17, 19, 23, 28
- Location of 30th percentile:

$$i = \frac{30}{100}(8) = 2.4$$

- The location index, i , is not a whole number; $i+1 = 2.4+1=3.4$; the whole number portion is 3; the 30th percentile is at the 3rd location of the array; the 30th percentile is 13.



25

Look at this example the raw data is given 14, 12, and 19 up to 17. I have arranged in the ascending order the lowest value is 5, the highest value is 28. Suppose I want to know 30th percentile for knowing the 30th percentile, first I have to find out i that is $(30 / 100)$

multiplied by 8 = 2.4. The i is nothing but location index as I explained the previous slide, i is not the whole number. So you have to add $i + 1$, so $2.4 + 1 = 3.4$.

In the 3.4 the whole number portion is 3 right? So the 30th percentile is at the 3rd location of an array. When you look at the 3rd location is 13, that means a person who scored 13 marks his corresponding percentile is 30.

(Refer Slide Time: 19:26)

Dispersion

- Measures of variability describe the spread or the dispersion of a set of data
- Reliability of measure of central tendency
- To compare dispersion of various samples



26

So far we talked about these different central tendencies will go for differing. Now we are going for measuring dispersion measures of variability describes the spread or the dispersion of the set of the data. The reliability of measure of central tendency is the dispersion because many times, the central tendency will mislead the people. So the reliability of that central tendency is calculated by or identified by its corresponding dispersion.

It is used to compare dispersion of various samples that is why whenever you plot the data you not only show the mean you have to show the central tendency also because the reliability of mean is explained by dispersion.

(Refer Slide Time: 20:14)

Variability



27

You look at this data, when you see the first two rows is no variability in cash flow mean is same. The second one is variability in cash flow see there is a lot of variability in the second one but the mean is same. If you look at only the mean it look like same when you look at only the mean the mean value same but when you look at see the left hand side the second dataset is having more variability. The quality of the mean is explained by its variability that is nothing but dispersion.

(Refer Slide Time: 20:51)

Measures of Variability or dispersion

Common Measures of Variability

- Range
- Inter-quartile range
- Mean Absolute Deviation
- Variance
- Standard Deviation
- Z scores
- Coefficient of Variation

28

There are different measures to measure the variability one is the range, inter-quartile range, mean, absolute deviation, variance, standard deviation, z scores and coefficient of variations. We will see one by one.

(Refer Slide Time: 21:07)

Range – ungrouped data

- The difference between the largest and the smallest values in a set of data
- Simple to compute
- Ignores all data points except the two extremes
- Example:

$$\text{Range} = \text{Largest} - \text{Smallest} = 48 - 35 = 13$$

35	41	44	45
37	41	44	46
37	43	44	46
39	43	44	46
40	43	44	46
40	43	45	48

29

Suppose there is ungroup of data is there see this one you have to find out the range. The range is nothing but the difference between the largest and the smallest value in a set of data. It is very simple to compute. The problem here is it ignores all data points except the two extremes. So the range is the largest value is 48 in this data set the smallest value is 35. $48 - 35 = 13$ you see that only the two values are taken care in between the values is not taken into consideration for finding the range.

(Refer Slide Time: 21:46)

Quartiles

- Measures of central tendency that divide a group of data into four subgroups
- Q_1 : 25% of the data set is below the first quartile
- Q_2 : 50% of the data set is below the second quartile
- Q_3 : 75% of the data set is below the third quartile
- Q_1 is equal to the 25th percentile
- Q_2 is located at 50th percentile and equals the median
- Q_3 is equal to the 75th percentile
- Quartile values are not necessarily members of the data set

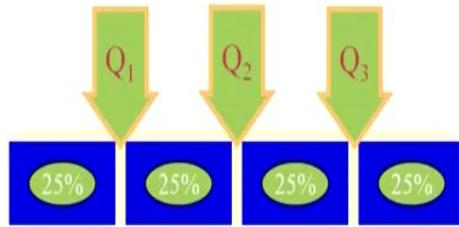
30

It is a quick estimate to measure the dispersion of a set of data. I will go for a quartile; quartile measures the central tendency that divided group of data into 4 subgroups. We say Q 1, Q 2, Q 3. Q 1 is nothing but 25 % of the data set is below the first quartile. Q2, 50 % of the data set is below the second quartile. Q3, 75 % the data set is below the third quartile. So we can say Q 1 is the 25th percentile Q 2 is the 50th the percentile nothing but the median.

This is a very important point; Q2 is nothing but the median. Q3 is the 75th percentile and another point is the quartile values are not necessarily members of the data set.

(Refer Slide Time: 22:34)

Quartiles



31

You see this lets say Q1, Q2, Q3. So Q1 see first 25 % the data set, Q2 first 50 % of the data set Q3, 75 % of the data set Okay? It is nothing but the quartile is used to divide the whole data set into 4 groups first 25, second 25, third 25 and last 25.

(Refer Slide Time: 22:59)

Quartiles: Example

- Ordered array: 106, 109, 114, 116, 121, 122, 125, 129

- Q_1 $i = \frac{25}{100}(8) = 2$ $Q_1 = \frac{109 + 114}{2} = 111.5$

- Q_2 :

$$i = \frac{50}{100}(8) = 4 \quad Q_2 = \frac{116 + 121}{2} = 118.5$$

- Q_3 :

$$i = \frac{75}{100}(8) = 6 \quad Q_3 = \frac{122 + 125}{2} = 123.5$$

32

Suppose an example for finding the quartile, suppose the data is given 106, 109 and so on. Okay? We have arranged it in the ascending order. First we got to find out the Q 1, Q 1 as I told you but the 25th percentile so the location of the 25th percentile. First you have to find out the location index i for the 25/ 100 x 8 = 2. Since the 2 is the even number. As I explained

previously if it is the location takes it 2 you have to find out that position plus the next position and its average.

So in the second positions data set is $109 + 114 / 2 = 111.5$. So the Q1 is nothing but here 111.5, Q 2 is 50th percentile $50 / 100 \times 8 = 4$, again the 4 is the even number. So the 4th location is 1, 2, 3, 4th location is 116 and 5th the location is 121 so, $116+121 / 2 = 118.5$. So the Q2 our median is 118.5. Then Q3, $75 / 100 \times 8 = 6$, 6 is the even number the average of 6th and 7th values are $122 + 125 / 2. = 123.5$.

(Refer Slide Time: 24:28)

Interquartile Range

- Range of values between the first and third quartiles
- Range of the “middle half”
- Less influenced by extremes

$$\text{Interquartile Range} = Q_3 - Q_1$$



33

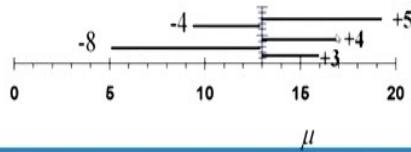
This is the way to calculate Q 1, Q 2, and Q 3. Now the next term is interquartile range .So the dispersions in the data set is measured with help of interquartile range by using this formula $Q_3 - Q_1$. As we know Q 3 is 75th percentile Q1 is the 25th percentile so range of values between the first and third quartile is called interquartile range. It is a range of middle of .Why we are using quartile range because it is the less influenced by extreme values.

Because when we collect the data set we are not going to consider at very low values at the same time very high values. So the middle values which is not affected by extremes that is taken for further calculation .For that purpose we are using interquartile range.

(Refer Slide Time: 25:15)

Deviation from the Mean

- Data set: 5, 9, 16, 17, 18
- Mean: $\mu = \frac{\sum X}{N} = \frac{65}{5} = 13$
- Deviations from the mean: -8, -4, 3, 4, 5



34

There is a Q3, now we will go for deviation from the mean so dataset is the given 5, 9, 16, 17, and 18. To find the deviation from the mean first to find the mean, mean is 13. Suppose there is a graph is there so this is the 13 Okay? See the first value 5 the difference is $5 - 13 = -8$. So this distance is your first deviation the second data is 9. So $9 - 13 = -4$ this is -4. So this deviation is expressed by these lines.

Look at that there is a negative deviation there is a positive deviation. Suppose if we want to add the deviation general it will become 0. That is why we should go for mean absolute deviation.

(Refer Slide Time: 26:12)

Mean Absolute Deviation

- Average of the absolute deviations from the mean

X	$ X - \mu $	$ X - \mu $
5	-8	+8
9	-4	+4
16	+3	+3
17	+4	+4
18	+5	+5
	0	24

$$M.A.D. = \frac{\sum |X - \mu|}{N}$$

$$= \frac{24}{5}$$

$$= 4.8$$

35

You see this X is given here there are 2 values are negative deviation 3 are three values are positive deviation. When you add this it is becoming 0 so it seems we are getting 0 we cannot

measure the dispersion. One way is we have to remove this negatives you take only positive value. When you take positive values $24 \text{ so } = 24 / 5$ there are 5 data set, $= 4.8$ is called mean absolute deviation. It is the average of absolute deviations from the mean.

(Refer Slide Time: 26:46)

Population Variance

- Average of the squared deviations from the arithmetic mean

X	$X - \mu$	$(X - \mu)^2$
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
	0	130

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

$$= \frac{130}{5}$$

$$= 26.0$$


36

There was a problem in the mean absolute deviation I will tell you what is the problem there, see the next we will see population variance it is not the average of the squared deviation from the arithmetic mean. Okay? So the X is there, mean is there so when you add the absolute even digit is 0, one way the previously the mean absolute deviation you take an only positive value. Now we are going to square it, the squaring of the deviation having some advantage.

One advantage is we can remove the negative sign second one is but the deviation is less when you square it. For example; - 4 square is 16 see - 8 squared is 64. So what is happening

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

more the deviation more this squared value. Okay? So

$$= 130/5$$

here we are squaring the purpose why we are squaring for there are two reason one is to remove the negative sign, the second reason is giving higher penalty for higher deviation values

The next one is the population standard deviation because already there is a variance but variance is a squared number that we cannot compare. Suppose the two numbers are given say 12 and 13 that is easy intuitively we can say which is higher which is smaller. Suppose 124, 169 is given notice squared number. We cannot compare intuitively and not only that it is in the square root of squared term.

We want to have it in the actual term so for comparison purpose for that purpose we are taking square root of that.

(Refer Slide Time: 28:25)

Population Standard Deviation

- Square root of the variance

X	$X - \mu$	$(X - \mu)^2$
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
	0	130

$$\begin{aligned}\sigma^2 &= \frac{\sum (X - \mu)^2}{N} \\ &= \frac{130}{5} \\ &= 26.0 \\ \sigma &= \sqrt{\sigma^2} \\ &= \sqrt{26.0} \\ &= 5.1\end{aligned}$$



37

So 5.1 is the standard deviation next we will go to the sample variance the formula is same but only thing is it is divided by $n - 1$.

(Refer Slide Time: 28:37)

Sample Variance

- Average of the squared deviations from the arithmetic mean

X	$X - \bar{X}$	$(X - \bar{X})^2$
2,398	625	390,625
1,844	71	5,041
1,539	-234	54,756
1,311	-462	213,444
7,092	0	663,866

$$\begin{aligned}S^2 &= \frac{\sum (X - \bar{X})^2}{n-1} \\ &= \frac{663,866}{3} \\ &= 221,288.67\end{aligned}$$



38

Why we are dividing by $n - 1$, the reason is that to make the variance as the unbiased estimator. This is due to degrees of freedom since we already know the value of the mean will last one degrees of freedom. That we are dividing by $n - 1$ so it is very important whenever you find the sample variance so the in the denominator there should be a $n - 1$. So here the variance is 221,288.67.

(Refer Slide Time: 29:06)

Sample Standard Deviation

- Square root of the sample variance

X	$X - \bar{X}$	$(X - \bar{X})^2$
2,398	625	390,625
1,844	71	5,041
1,539	-234	54,756
1,311	-462	213,444
7,092	0	663,866

$$S^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

$$= \frac{663,866}{3}$$

$$= 221,288.67$$

$$S = \sqrt{S^2}$$

$$= \sqrt{221,288.67}$$

$$= 470.41$$

This is another sample standard deviation; just to take the square root of that it is a 470.41 so a square root of the variance is nothing but standard deviation.

(Refer Slide Time: 29:17)

Uses of Standard Deviation

- Indicator of financial risk
- Quality Control
 - construction of quality control charts
 - process capability studies
- Comparing populations
 - household incomes in two cities
 - employee absenteeism at two plants

Now the purpose is why we have to study the standard deviation because the standard deviation is giving an indicator of financial risk. Higher the standard deviation is more risk lesser the standard deviation less at risk. In quality control context generally when we

manufacture something suppose here plant A and plant B or shift A and shift B whenever the variances in lesser then the quality of the product is high.

The process capability also they should have the lesser variance means in the process capabilities high. Then I suppose therefore comparing the populations household income of 2 cities, employee absenteeism in 2 plants for these purposes, it is for comparing the population that means wherever there is a lesser standard deviation so that is having higher homogeneous data set.

(Refer Slide Time: 30:12)

Standard Deviation as an Indicator of Financial Risk

Annualized Rate of Return		
Financial Security	μ	σ
A	15%	3%
B	15%	7%

41

You see look at this one μ and σ , see this is a financial security A and B. See the return rate is 15, 15 it is both are giving equal return but look at this the σ standard deviation because in financial context it is, it is measured as the risk. So the first one is 3% second with 7% so the security B having a higher risk, so always we will go for where there is a lesser standard deviation because mean is same.

We are the same time the risk all should be same.

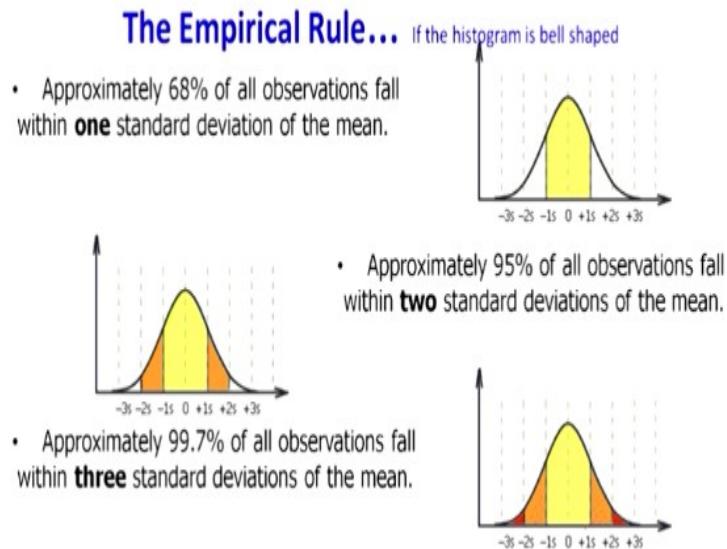
So far we have seen different central tendencies, different dispersions. In the coming class will use Python will take some sample data set. I will explain you how to find out central tendency and the dispersion of the given data set. Thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology, Roorkee

Lecture No 5
Central Tendency and Dispersion

Good morning students; today we are going to the lecture number 5, Central tendency and dispersion will continue what have stopped, from the previous lecture, what we are going to see today is, one important property of a normal distribution. And second, we will see various kurtosis. Then second we see the box and whisker plots that has a different way of measuring the dispersion.

(Refer Slide Time 00:51)



See this empirical rule, if the histogram is Bell Shaped. Look at this normal distribution, Bell shaped curve. This yellow line says that, from the mean if you are traveling either side on 1 sigma distance. You can cover 68 % of all observations. Come to the second one, from the means 0. If, you travelling 2 sigma distance on the either side. You can cover 95% of all observations.

The third one, if you are travel 3 sigma distance on either side from the mean of a normal distribution. You can cover 99.7 % of all observations. This is very important empirical rule.

Even through you study in detail about the normal distribution and it is the properties in the coming lectures, I wanted to say this idea may be very useful in coming lectures, because we can say the normal distribution is the father of all the distributions.

Because if you have any doubt on nature of the distributions. If you are not really sure about what distribution the data follow, you can assume normal distribution. But there is a limitation of this Empirical Rule. It is applicable only for the Bell shaped curves. There may be a situation, the actual phenomena need not follow bell shaped curve. At that time this formula that is the empirical rule will not work. So we will go to another rule, in the next slides, the same thing.

(Refer Slide Time 02:24)

Empirical Rule

- Data are normally distributed (or approximately normal)

Distance from the Mean	Percentage of Values Falling Within Distance
$\mu \pm 1\sigma$	68
$\mu \pm 2\sigma$	95
$\mu \pm 3\sigma$	99.7

What I given the previous slide, see $\mu \pm \sigma$. You can cover 68% of the all observations. $\mu \pm 2\sigma$, you can cover 95 % of the observations. $\mu \pm 3\sigma$, you can cover 99.7 % of all observations. Actually this 1, 2, 3 is nothing but Z. I will tell you incoming classes, what is Z.

(Refer Slide Time 02:57)

Chebyshev's Theorem

...Not often used because interval is very wide.

- A more general interpretation of the standard deviation is derived from **Chebyshev's Theorem**, which applies to all shapes of histograms (not just bell shaped).
- The proportion of observations in any sample that lie within k standard deviations of the mean is *at least*:

$$1 - \frac{1}{k^2} \text{ for } k > 1$$

For $k=2$ (say), the theorem states that at least 3/4 of all observations lie within 2 standard deviations of the mean. This is a "lower bound" compared to Empirical Rule's approximation (95%).

The previously we have seen that the properties of normal distribution, that is a bell shaped curve. Sometimes certain phenomenon need not follow the bell shaped curve. That time, you cannot use that property which he studied previously; you had to go for another formula for to find out how much observations are covering under 1σ , 2σ and 3σ distance. This idea was given by Chebyshev's. It is called Chebyshev's theorem.

Yeah, more general interpretation of the standard deviation is derived from Chebyshev's theorem, which applies to all shape of the histogram, not just to bell shape. Previously we see, what is totally for bell shaped, but here it is apply it is applicable to all shape of the histogram, even the distribution can follow any shape. So the proportion of observation in any sample that lie within k standard deviation of the mean is at least: $1 - (1/k^2)$.

So, how to read this formula is, suppose a phenomenon which is not following normal distribution. If you want to know the 2σ distances, how much percentage of observation can be covered? So when you substitute here $1 - (1/k^2) = 1 - 1/4 = 3/4$. So, 3/4 means 75 %. So if you are travel 2σ distance on the either side. For a distribution which is not following normal distribution. You can cover 75 % of all observations.

You see the previously; It is 95 %. So you see that that is a given. For k equal to 2, the theorem states that, at least 3/4th of all observations lie within two standard deviation of the mean. This is

lower bound compared to empirical rule approximation 95 %. In case the previous slide, if it is 2. We can cover 95% of all observations, but here we can cover only 75% of all observations. Sometime we can use Chebysheff's theorem also; the data is not following normal distribution.

(Refer Slide Time 05:14)

Coefficient of Variation

- Ratio of the standard deviation to the mean, expressed as a percentage
- Measurement of relative dispersion

$$CV = \frac{\sigma}{\mu} (100)$$

These two properties in coming classes many times we will refer that 1 sigma, 2 sigma, 3 sigma. Just like that I want to give an idea about the normal distribution, but we will go in detail later. The next way to measure the dispersion is coefficient of variation. The ratio of standard deviation to the mean, express as a percentage. So coefficient of variation is your sigma by mu, it is the measurement of relative dispersion. Already there is a standard deviation is there.

What is the purpose of this coefficient of variations that will see the next slide?

(Refer Slide Time 05:53)

Coefficient of Variation

$$\mu_1 = 29$$

$$\sigma_1 = 4.6$$

$$C.V_{.1} = \frac{\sigma_1}{\mu_1} (100)$$
$$= \frac{4.6}{29} (100)$$
$$= 15.86$$

$$\mu_2 = 84$$

$$\sigma_2 = 10$$

$$C.V_{.2} = \frac{\sigma_2}{\mu_2} (100)$$
$$= \frac{10}{84} (100)$$
$$= 11.90$$

Look at this, you see, stock A, stock B or stock 1, stock 2 the $\mu_1 = 29$, $\sigma_1 = 4.6$. Another one, $\mu_2 = 84$, $\sigma_2 = 10$, supposed to choose which is better. If you compare only the mean, 29 versus 84, the second stock is better. Suppose if you compare the standard deviation, 4.6 and 10. The lower the standard deviation better it is. So the stock 1 is better. Now there is a contradiction, with respect to mean option B is better with represent standard deviation option A is better.

Now there is a contradiction to need to have the trade off. In this situation we have to go for this coefficient of variation, coefficient of variation = σ / μ . For example, for this case; $= (4.6 / 29) \times 100 = 15.86$, for second case $= \sigma_2 / \mu_2 = (10 / 84) \times 100 = 11.90$. Lower the coefficient of variation, but have the option is. So, if the variance is smaller, to be able to choose that group, or that stock.

(Refer Slide Time 07:18)

Variance and Standard Deviation of Grouped Data

Population	Sample
$\sigma^2 = \frac{\sum f(M - \mu)^2}{N}$	$S^2 = \frac{\sum f(M - \bar{X})^2}{n - 1}$
$\sigma = \sqrt{\sigma^2}$	$S = \sqrt{S^2}$

Now we will see how to find the variance and standard deviation of the grouped data. Already we have seen the standard deviation of the raw data also named as ungrouped data. We are seeing the standard deviation for population, standard deviation for sample. Similarly, we will

$\sigma^2 = \frac{\sum f(M - \mu)^2}{N}$
find the variance for a grouped data. So the formula is:

Here, f is the frequency, M is the midpoint of that interval, mu is the mean of the interval, N is some of all frequencies.

$\sigma^2 = \frac{\sum f(M - \bar{X})^2}{n - 1}$
For the sample variants, look at the formula here it is also n-1,

(Refer Slide Time 08:03)

Population Variance and Standard Deviation of Grouped Data($\mu=43$)

Class Interval	f	M	fM	$M - \mu$	$(M - \mu)^2$	$f(M - \mu)^2$
20-under 30	6	25	150	-18	324	1944
30-under 40	18	35	630	-8	64	1152
40-under 50	11	45	495	2	4	44
50-under 60	11	55	605	12	144	1584
60-under 70	3	65	195	22	484	1452
70-under 80	1	75	75	32	1024	1024
	50		2150	/50		7200

$$\sigma^2 = \frac{\sum f(M - \mu)^2}{N} = \frac{7200}{50} = 144$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{144} = 12$$

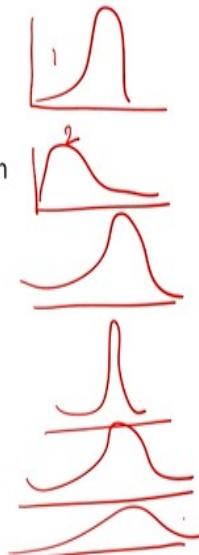
So we will see this example now look at this, there with the grouped data is given; but the problem is given in the form of table. So, see these are the interval. These are frequencies. So 20 to 30, there are six values are there. Between 30 and 40 there are 18 values is there, 40 and 50, 11 is there 50 and 60, 11 is there. 60 and 70, 3 is there, and so on. So first 1 we are to find out the midpoint of the interval between 20 and 30, the midpoint is 25, 35, Next interval 45, 55, 65, 75. Now, you multiply this f and the midpoint of the interval. So, 150, 630, and so on.

We need this $\sum fM$ then you can find $M - \mu$, M is 25- 43- 18, 35- 43-8 and so on. Then you square it, then the squared value you multiply by f , you are getting this $f(M - \bar{X})^2$. When you add it, that is going to be 7200, $N = \sum f$. There is nothing but 50. So 144 is the population variants of this grouped data. If you want to know the standard deviation of this group of data, just to take the square root of the variance, that is 12.

(Refer Slide Time 09:30)

Measures of Shape

- **Skewness**
 - Absence of symmetry
 - Extreme values in one side of a distribution
- **Kurtosis**
 - Peakedness of a distribution
 - Leptokurtic: high and thin
 - Mesokurtic: normal shape
 - Platykurtic: flat and spread out
- **Box and Whisker Plots**
 - Graphic display of a distribution
 - Reveals skewness



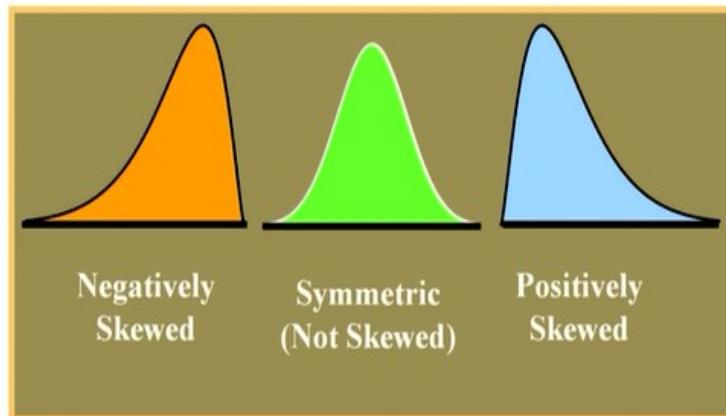
The next measure is shape of the, we can say a set of data. That is shape or distribution, what distribution follows. We can see the skewness, kurtosis, box and whisker plots there are the three method. So we will see what is a skewness, skewness is the absence of symmetry. As I told you it may be this is left skewed data. This is a right skewed data. This is symmetric data. So this absence of symmetry, this and this can be done with helps skewness.

So the other one the application of skewness is to find out what is the nature of this shape, whether it is skewed or symmetric. Next one is a kurtosis; it is the peakedness of a distribution. There are three layers, there are Leptokurtic, Mesokurtic, Platykurtic. Leptokurtic means high and thin, Mesokurtic is little flat in this way and Platykurtic very very very flat this way flat and spread out.

Then we can see, box and whisker plots. It is a graphical display of distribution. It reveals skewness. The application of boxer whisker plot is to check whether the data, follow a symmetry or what is the nature of the skewness of the distribution.

(Refer Slide Time 10:58)

Skewness



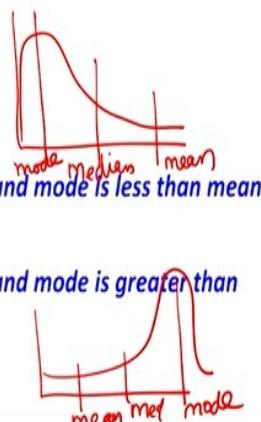
See the skewness left one which is an orange color it is the negatively skewed. As I told you the previous lecture skewness is how it is named is looking at the tail. the tail is on the left hand side so it is a left skewed or negatively skewed. Come to the blue one, the extreme right. It is this tail is on the right hand side. So it is a right skewed or positively skewed one, the middle one there is no skewness, so it is symmetric.

(Refer Slide Time 11:29)

Skewness..

The skewness of a distribution is measured by comparing the relative positions of the mean, median and mode.

- Distribution is *symmetrical*
 - *Mean = Median = Mode*
- Distribution *skewed right*
 - *Median lies between mode and mean, and mode is less than mean*
- Distribution *skewed left*
 - *Median lies between mode and mean, and mode is greater than mean*



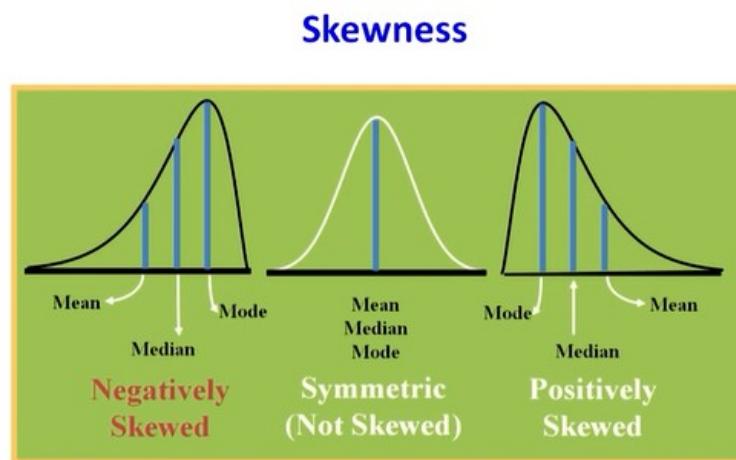
The skewness of a distribution is measured by comparing the relative position of the mean, median, and mode. If the distribution is symmetrical, we can say mean equal to median equal to mode. The distribution is skewed right, the median lies between mode and mean, and the mode

is less than mean. The distribution skewed right means this way. So this will be our mean, this will be our median, this will be mode.

Look at this, Median lies between mode and mean. The mode is less than the mean because mode is less than the mean. The same thing the distribution is skewed left. This is the case. So the mean position of mean will be here. Median position of the mode, the median is lies between mode and mean. And the mode is greater than mean. As I told you, whenever, if you want to know the central tendency of your distribution you to check the nature of the distribution.

If it is skewed right or left, you should go for median, because the median always middle of the distribution irrespective of your skewness.

(Refer Slide Time 12:58)



The same thing, what are explained the previous one. Mean, see negatively skewed the position of mean is here, median, mode. Positively skewed to the position of means here, median is here, mode is here. There is no skewness in middle one.

(Refer Slide Time 13:11)

Coefficient of Skewness

- Summary measure for skewness

$$S = \frac{3(\mu - M_d)}{\sigma}$$

- If $S < 0$, the distribution is negatively skewed (skewed to the left)
- If $S = 0$, the distribution is symmetric (not skewed)
- If $S > 0$, the distribution is positively skewed (skewed to the right)

How to find coefficient of Skewness? The summary measure for Skewness can be measured as:

$$S = \frac{3(\mu - Md)}{\sigma}$$

If S is negative, it is negatively skewed. If S equal to 0. It is symmetric. If S is greater than 0, the distribution is positively skewed.

(Refer Slide Time 13:35)

Coefficient of Skewness

$\mu_1 = 23$	$\mu_2 = 26$	$\mu_3 = 29$
$M_{d1} = 26$	$M_{d2} = 26$	$M_{d3} = 26$
$\sigma_1 = 12.3$	$\sigma_2 = 12.3$	$\sigma_3 = 12.3$
$S_1 = \frac{3(\mu_1 - M_{d1})}{\sigma_1}$	$S_2 = \frac{3(\mu_2 - M_{d2})}{\sigma_2}$	$S_3 = \frac{3(\mu_3 - M_{d3})}{\sigma_3}$
$= \frac{3(23 - 26)}{12.3}$	$= \frac{3(26 - 26)}{12.3}$	$= \frac{3(29 - 26)}{12.3}$
$= -0.73$	$= 0$	$= +0.73$

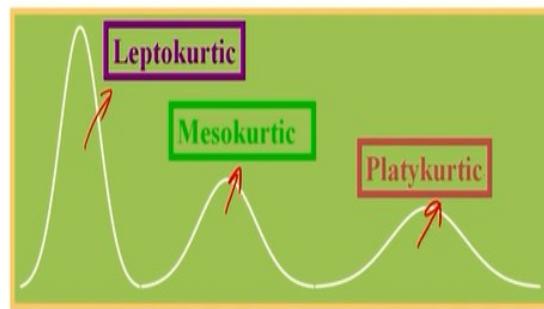
You will see an example. μ_1 is 23, median1 is 26, σ_1 is 12.3, and you apply this formula, $= 3 \times (23 - 26) / 12.3$ we are getting negative, so it is a negatively skewed. Go to the middle one μ_2 equal to 26, median2 equal to 26, so $26 - 26 = 0$. So S_2 equal to 0. For this distribution the

skewness is 0 or it is symmetric. The right one μ_3 equal to 29, median is 26, σ_3 is 12.3, and you substitute here we are getting positive value for S_3 . So the skewness is positive.

(Refer Slide Time 14:20)

Kurtosis

- Peakedness of a distribution
 - Leptokurtic: high and thin
 - Mesokurtic: normal in shape
 - Platykurtic: flat and spread out



The kurtosis, as I told you, it is a peakedness of a distribution, when they say Leptokurtic, leptokurtic, this one. So it is high and thin, if that means highly homogeneous distribution, the things are very close. This is second one is the Mesokurtic, it is normal shape. The last one is Platykurtic, flat and spread out.

(Refer Slide Time 14:48)

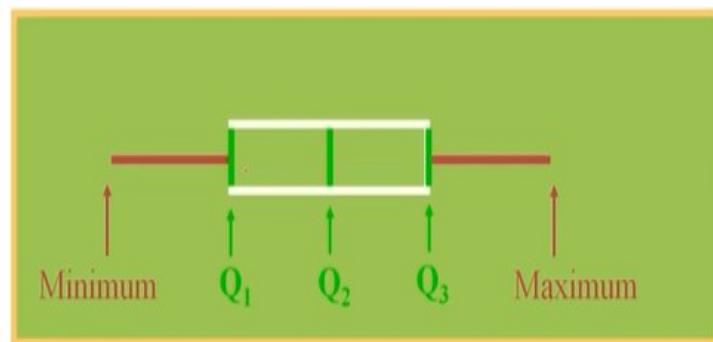
Box and Whisker Plot

- Five specific values are used:
 - Median, Q_2
 - First quartile, Q_1
 - Third quartile, Q_3
 - Minimum value in the data set
 - Maximum value in the data set

The next we will go to box and whisker plot. There are five positions in the box and whisker plot. One is median Q2, first quartile Q1, third quartile Q3. The next word is the minimum value in the data set, maximum value in the data set.

(Refer Slide Time 15:05)

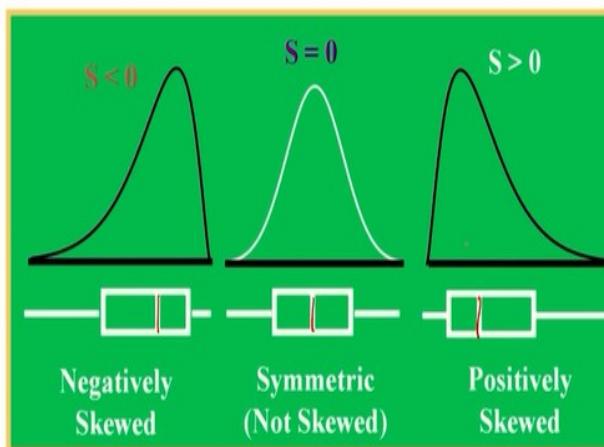
Box and Whisker Plot



We will see this one here, this one. So, this point is would minimum value in this box is a Q1 is the quartile one, quartile two, quartile three maximum, why its called box and whisker plot. The whisker is look like a whisker of a cat. So it is a box and whisker plot.

(Refer Slide Time 15:25)

Skewness: Box and Whisker Plots, and Coefficient of Skewness



You see how the skewness can be measured or identified with the help of box and whisker plot. So by looking at the position of this middle this line, we can identify the distribution. What is

that, if it is on the right side of this box it is left skewed data. If it is a left side it is the right skewed data. If it is exactly on the middle, it is symmetric which follow normal distribution. So far, we have given some kind of theories about this various central tendencies and dispersion.

Now, I'm going to switch over to Python. So whatever we have done. The theory portion whatever we are taught here, so I am going to use Python. I am going to explain how to use Python to get central tendencies, skewness, box and whisker plot and various dispersion techniques with the help of Python. So we will go to the Python mode.

(Video Starts: 16:26)

Okay, now we will come to the Python environment, the first, as I told you we are to import pandas as pd. We can do this pd is it is for only our convenient fantasies and library. Then they are to the import Numpy numerical Python, as the np. Okay, so the first one is to import the required libraries. The second one is going to import the data set. So the data set, already I know the part of the data set is the otherwise the name of the data set is IBM underscore 313 marks.xlsx.

So, I am going to save the object called Table. Table equal to pd., this is the command. This read_excel is the command for the reading the Excel file. The path is this, 'IBM-313' otherwise; simply you can type it there. Now print table, let us see what is the data. See, look at this, serial number is there, MTE that is a midterm examination marks, mini project, total, end term examination marks and total marks.

Okay, this total is out of 100, this total is out of 50. You see it is starting from 0,1,2,3. Okay, now I want to find out the mean of the total that is in the end term examinations. So, x = table, the object either be a square bracket. There you have to write the column name. So that means I am going to take only the column name, total, and I go to store that value into the variable x. Now the x is nothing but by the last column, if you want to mean of that one. So, np.mean.

Otherwise, if you want to know np. If you press tab you will get various options to that np. tab. See here, there are so many options there in that have there are maximum, minimum, the mean, median, you need not remember also you can check it one by one. So now we will go to the

`np.mean`. Then, `np.mean`, then we will call that variable `x` executed, shift enter. So we are getting this value 46.90 is the average marks.

There is a lot of the median. So `np.median`, median is 45. We will go for mode, the mode. You have to import `scipy`, from `scipy import stats`. Stat is another library function. So start `stats.mode` called the variable `x`. We will see what is the mode? Mode is the number of frequencies. Suppose, there are five students see got the same marks 30 bars, the mode will be 30. Okay. So, okay, we will come back to later.

So, next we will go to percentile. In percentile suppose I have taken the array, that we are introducing another one `np.array`, `a` equal to `np.array` just we have taken an array 1, 2, 3, 4, 5. Suppose, go to say `p` equal to `np.percentile` of that array `a`, 50. What do you want to know? I want to know 50th percentile, 50th percentile means what value in this array will be the 50th percentile, and execute this print `P`. So, 3 is the 50th personality but the median.

This number is very small number one for illustration purpose. So you can have a large number then you can run it. Then now, we will go to another command in Python is for loop. For loop is suppose I have taken a variable `k` saved three variables, one is Ram, Seeing the characters in the code 65, 2.5. Suppose if I print `k`, what will happen. You see that it is printing Ram, 65, 2.5. But there is a requirement that I have to print one by one.

First I have to print Ram, and then I have to print 65, and then print 2.5. Now here at a time I am getting all the answer but I want to print one by one. So for that purpose, see that for `i` in `k`. This is the syntax, there should be a colon, that is one print `i`. So what will happen first in `k` this is array. So first, for `i` in `k` will take the value Ram, second `i` will take the value of 65. Third, `i` will take the value of 2.5.

Now if we execute this print `i`, see that one by one. So this is the one by one, I am getting this output. So this is the example of for loop. So for `i` in `k`. the `k` is in which variable. So the `i` value will change. if you want to print `i`. So the first it print Ram and then 65 and 2.5, because why I

am showing that we are going to use this for loop incoming examples. So I want to give an idea about how to use for loop in Python.

Now we will go to the range. So, far i in range. Is it that 10, 20, 2 this is the range function. The range first one is the starting value. The second one is the ending value, the 2 is increment; print i. if we print that you see that, now what is happening. 10, 12, 14, 16, 18 incremented by 2, ending with excluding 20. 1, 2, 3, 4, 5 increment is by 2. Now suppose in the print, I want to print, now it is printed one by one.

But I want to print in i with the comma, so 10, 12, 14 that purpose the same comment, i use end equal to it should be separated by a comma in colon. So if I run this. What is happening see 10, 12, 14, 16, 18 so this one end equal to in ','. That is what how it is giving the output in horizontal way. Now we will go to the next option functions in Python. Suppose the functions are very useful applications in Python many time.

There are some built in function is there. For example; print is the built in function, maximum is the built in function, and minimum is built in function. You can create your own functions, and then you can call that function wherever it is required. Suppose def, that is the syntax def greet open parenthesis, end with the colon, print Hi, print good evening. So, this is the way of defining your function.

Okay, then. After defining the function, you have to call that function, suppose I call the greet. I will execute this word what will happen? So, this function is getting executed. So again Hi, and good evening. So another example or function suppose I want to add two numbers, so def the function name is add in parenthesis, p,q. It can be anything colon, the colon is important. Otherwise, it will show syntax error.

So c equal to p+q, so print c. suppose this is my function, suppose this is, I want to call this function, add 6,4 what answer I am getting. Suppose other number suppose, add 10,4. So I am getting 14. We have seen how to create a function now finding the minimum, maximum value in

the data set. Suppose I created a new array. Data equal to 1, 3, 4, 463, 2, 3, 6, just i take randomly. Suppose I want to see the minimum value in this array and maximum value in this array.

So what is happening. So minimum value is 1, the maximum value is 463. So for that the comment is min and max. Now, this minimum and maximum value, I can create own function. Then I can call that function. Because every time I need not type minimum, min data, max data. Because already that was built in function, we can create our own function. So the same data I have taken, 1, 3, 4, and 463.

I am defining function min underscore; underscore MAX data, so min underscore value equal to minimum data, maximum underscore value equal to maximum data. Now returns because I want to get the output. This is indentation is more important. Suppose, there is, minimum_value should be the same, same indentation, generally we can give Tab. Tab means we can save for space work. So return underscore, minimum_value, maximum_value will run it let us see what is happening.

So I am getting this because I called this function again how? Min_and_max(data). Because my function name is Min_and_max. So 1, 463. So this functions application is very much useful in Python because when you are making a large program, every time some routine aspects you need not do it, yourself every time. So you can call that function whenever it is required. It will save a lot of your time and energy.

So now, suppose I want to know the range of the data range is nothing but maximum value and minimum value. For that, I go to define a function def, that function name is rangef. That is the rangef. rangef I given you can give any name for the data. So, I am finding

```
minimum_value = min(data)
maximum_value = max(data)
return (maximum_value - minimum_value)
```

So if i call that function rangef data, what will happenning I getting 462, nothing but = 463-1. Now we will go to quartile. Quartile, we have seen already. It is a Q1, Q2, Q3. Q1is the 25th

percentile, there is an inbuilt function in NumPy. So when you say I am creating an array, a equal to np.array, array1, 2,3,4,5. So Q1 equal np.percentile. This np.percentile will give you the percentile.

Suppose if you want to know 25th percentile I can get what value in this array is the 25th percentile. So if I execute this. So, that means, the value in two is 25th percentile. So the same thing np.percentile(a,50), if I put. I am going to call it Q2. So 3 is, otherwise median you see look at this because it is odd number, the middle value three obviously it is the 50th percentile. I will go for third one, Q3 = 4. That is our 75th percentile.

Next, spredness is measured in terms of Inter quartile range. As we know already, Q3-Q1, so that is nothing but IQ. So if we see IQ, there is nothing but Q3-Q1, 2 is inter quartile range. Now, we will go for how to find out the variance. Suppose, there are two way two variants one variances for fine variants of mean, another one is the variance of the population. Even you put np. NumPy, var is for the population variance x. So what is the x, x is the total.

So that column, the total column we have saved in the name of object called x or variable called x. So we will see variance is 262.781. Suppose, there is a another function will be import to library statistics. So statistics.pstdev, that is population standard deviation x. so I can get the population standard deviation, that is 16.2105. It is the standard deviation. So if you want to know the standard deviation of the sample. So, statistics.stdev.

So only thing is, if you want to know the population standard deviation, you should write p there, np.std otherwise by default, in statistics library, you are getting only the sample standard deviation. This is standard deviation for the population. This is standard deviation for the sample. Next round for the skewness for that you are to import from scipy there is a library called dot stats import skew, skewness of x. So skewness is positive value so it is the right skewed data.

Next we will go to Box and whisker plot, because for drawing plot you have to use Matplotlib. Import pyplot as plt. So plt.boxplot, there is an inbuild function as x comma symbol is star (*) plt.show execute this. So we are getting box and whisker plot. You see that box and whisker plot

rather than some star symbol that implies that that data is are outlier. Outlier means which goes beyond maximum value beyond minimum value.

The position of this middle line will help you to identify the nature of the distribution. If it is in left side, it is right skewed data. See, look at this, because it should positively skewed data and now it is little left side. So the data is, data is right skewed data. So with that we are stopping the central tendency. So what we have seen so far we have seen various central tendencies and different way of measuring the dispersions,

(Video Ends: 31:40)

Whatever you will learn theory part that we run in Python, we got the answer. There are so many sources are available in internet to know the different course, different videos find out you can also refer that for this class. Thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 06
Introduction to Probability - I

Good morning students. Today we are going to next lecture number 6 introduction to probability. The concept of probability is fundamental in any field whether you call it a statistics or analytics are your data science everywhere because if you look at some of the book titles of the statistics or analytics it will come with probability and statistics because the concept of probability and statistics is cannot separate it because always will go together because, the concept of statistics since we are taking sampling, we are predicting about the population.

So, whatever we say with the help of sampling we have to attach always some probability because it cannot be 100% assured that whatever you say with the help of sample will be exactly; you cannot predicted so when there is a prediction comes, then you have to attached probability to that. Today, you will see that it is an introduction to probability I am not going to teach in detail about that one.

(Refer Slide Time: 01:21) correct the picture size

Lecture objectives

- Comprehend the different ways of assigning probability
- Understand and apply marginal, union, joint, and conditional probabilities
- Solve problems using the laws of probability including the laws of addition, multiplication and conditional probability
- Revise probabilities using Bayes' rule

What are the ideas which are important for us the only that I am going to teach. So the lecture objective is to comprehend the different way of assigning probability understand and apply in

marginal union joint and conditional probabilities and solving problem using laws of probability including law of addition, multiplication and conditional probability and using very important theorem that is the Bayes rule that to revise the probability at the end these are the my lecture objectives.

(Refer Slide Time: 01:51)

Probability

- Probability is the numerical measure of the likelihood that an event will occur.
- The probability of any event must be between 0 and 1, inclusively
 - $0 \leq P(A) \leq 1$ for any event A.
- The sum of the probabilities of all mutually exclusive and collectively exhaustive events is 1.
 - $P(A) + P(B) + P(C) = 1$
 - A, B, and C are mutually exclusive and collectively exhaustive

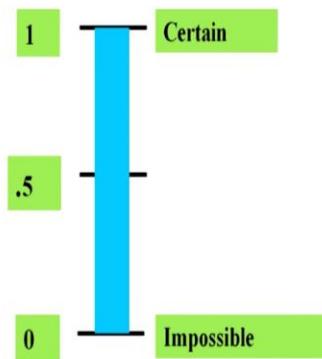


3

So we will go the definition of the probability, the probabilities the numerical measure of likelihood that an event will occur, the probability of any event must be been between 0 and 1, inclusively 0 to 1 for any event A, the sum of the probability of all mutually exclusive collectively exhaustive event is 1 latter will explain what would be mutually exclusive collectively exhaustive events. So always the probability of summation of probability will be equal to 1, for example, probability A plus probability B and probability C equal to 1, here A, B and C are mutually exclusive and collective events.

(Refer Slide Time: 02:32)

Range of Probability



4

So this is the range of probability you see that, if it is an impossible event, the probability value is 0. If it is a certain event, the probability is 1 if it is there a 50% chance, the probability is point 5. So, the point here is it is the probability lies between 0 to 1.

(Refer Slide Time: 02:50)

Methods of Assigning Probabilities

- Classical method of assigning probability (rules and laws)
- Relative frequency of occurrence (cumulated historical data)
- Subjective Probability (personal intuition or reasoning)



5

The method of assigning probability there are 3 methods, one is the classical method of assigning probability rules and laws, the relativity frequency of occurrence that is cumulative to historical data and subject to probability that is a personal intuition or reasoning by using these 3 methods, let us see how to find out the probability.

(Refer Slide Time: 03:11)

Classical Probability

- Number of outcomes leading to the event divided by the total number of outcomes possible
- Each outcome is equally likely
- Determined *a priori* -- before performing the experiment
- Applicable to games of chance
- Objective -- everyone correctly using the method assigns an identical probability



6

First of all is a classical probability the number of outcomes leading to an event divided by the total number of outcomes possible is a classical probability. Each outcome is equally likely there is an equal chance of getting different outcome. So, it is determined *a priori* that is before performing the experiment we know what are the outcome is going to come suppose we toss a coin there are 2 possibility head or tail in advance we know what are the possible outcomes.

It is applicable to games of chance the object to is everyone correctly using the method assign and identical probability because what is happening here, using classical probability that everyone will get the same answer for a problem because we know in advance what are the possible outcomes.

(Refer Slide Time: 04:00)

Classical Probability

$$P(E) = \frac{n_e}{N}$$

Where:

N = total number of outcomes

n_e = number of outcomes in E



7

So, mathematically the $P(E) = n_e / N$

where the N is the total number of outcomes n_e is the number of outcomes in E.

(Refer Slide Time: 04:15)

Relative Frequency Probability

- Based on historical data
- Computed after performing the experiment
- Number of times an event occurred divided by the number of trials
- Objective -- everyone correctly using the method assigns an identical probability



8

The relative frequency probability, it is based on the historical data because the another name for relatively frequencies of probability, it is computed after performing the experiment, number of items an event occurred divided by number of trials and it is nothing but frequency divided by sum of frequency, here also everyone correctly using this method assign as identical probability because everything is already known to you.

(Refer Slide Time: 04:42)

Relative Frequency Probability

$$P(E) = \frac{n_e}{N}$$

Where :

N = total number of trials

n_e = number of outcomes
producing E



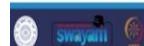
9

So, the same formula probability E is n_e divided by N where n_e is number of outcomes N is total number of trials.

(Refer Slide Time: 04:50)

Subjective Probability

- Comes from a person's intuition or reasoning
- Subjective -- different individuals may (correctly) assign different numeric probabilities to the same event
- Degree of belief
- Useful for unique (single-trial) experiments
 - New product introduction
 - Initial public offering of common stock
 - Site selection decisions
 - Sporting events



10

Then subject to probability, it comes from a person's intuitions or reasoning. Subjective means different individuals may correctly assign different numerical probabilities to the same event, it is the degree of belief sometimes subjective probability is useful for example, if you introduce a new product, suppose if you want to know the probability of success of the new product so we can ask an expert that what is the probability of success. Even if it is a new movie or new project what is the probability of success?

It is based on the intuition are based on the experience of the person he can give some probability of success or failure for example sites select decision for sporting events for example, in cricket what is the how much possibility of one team to win these are intuitive probability.

(Refer Slide Time: 05:42)

Probability - Terminology

- Experiment
- Event
- Elementary Events
- Sample Space
- Unions and Intersections
- Mutually Exclusive Events
- Independent Events
- Collectively Exhaustive Events
- Complementary Events



11

In this course that there are certain terminology with respect to probability you have to know it is very fundamental, even though you might have studied or even with previous classes it is just to recollect it what is the experiment we will see what is the experiment event, elementary event, sample space, union and intersections mutually exclusive events, independent events, collectively exhaustive events, complimentary events, these are some terms we will revise this.

(Refer Slide Time: 06:07)

Experiment, Trial, Elementary Event, Event

- **Experiment:** a process that produces outcomes
 - More than one possible outcome
 - Only one outcome per trial
- **Trial:** one repetition of the process
- **Elementary Event:** cannot be decomposed or broken down into other events
- **Event:** an outcome of an experiment
 - may be an elementary event, or
 - may be an aggregate of elementary events
 - usually represented by an uppercase letter, e.g., A, E1



12

One is we say what is experiment trial elementary event and event experiment, a process that produces outcomes is experiment so there are more than one possible outcome is there only one outcome per trial, what is a trial one repetition of process is a trial what is the elementary event? Even that cannot be decomposed or broken down into other events that is elementary events and what is the event and outcome of an experiment may be an elementary event maybe aggregate of elementary event usually represented by uppercase letter for example A, E that is notation for event.

(Refer Slide Time: 06:48)

An Example Experiment

- Experiment: randomly select, without replacement, two families from the residents of Tiny Town
- Elementary Event: the sample includes families A and C
- Event: each family in the sample has children in the household
- Event: the sample families own a total of four automobiles

Tiny Town Population		
Family	Children in Household	Number of Automobiles
A	Yes	3
B	Yes	2
C	No	1
D	Yes	2



13

We look at this example. If you look at the table there are some towns population is given there are 4 families their family A, B, C, D we asked the 2 questions, children's in household whether do you have children are not see family yes. Then we asked a number of automobiles how many number of automobiles you have 3. B they have children, they have 2 automobiles for the help of this table will try to understand what is experiment for example, randomly select without replacement 2 families from the residents of the town.

For example, randomly we can select so elementary event for example, the sample includes family A and C randomly you have to selected. So, event each family in the sample has children in the household, the sample families own a total of 4 automobiles to these are particular events for example for the event each family in the sample has children in the household for example A is one event D is another event for example, the sample families own a total number of 4

automobiles for example, A and C they have 4 automobiles B and D they have 4 automobiles A and D they have more than 4 automobiles.

(Refer Slide Time: 08:07)

Sample Space

- The set of all elementary events for an experiment
- Methods for describing a sample space
 - roster or listing
 - tree diagram
 - set builder notation
 - Venn diagram



14

This is the example of event then what is the sample space, the set of all elementary events for an experiment is called a sample space. Suppose if you roll a die, there are you can get 1 2 3 4 5 6 these are the sample space there are different methods for describing the sample space one is listing tree diagram, set builder notation and Venn diagram, you will see what is that.

(Refer Slide Time: 08:30)

Sample Space: Roster Example

- Experiment: randomly select, without replacement, two families from the residents of Tiny Town
- Each ordered pair in the sample space is an elementary event, for example -- (D,C)

Family	Children in Household	Number of Automobiles
A	Yes	3
B	Yes	2
C	No	1
D	Yes	2

Listing of Sample Space
(A,B), (A,C), (A,D), (B,A), (B,C), (B,D), (C,A), (C,B), (C,D), (D,A), (D,B), (D,C)



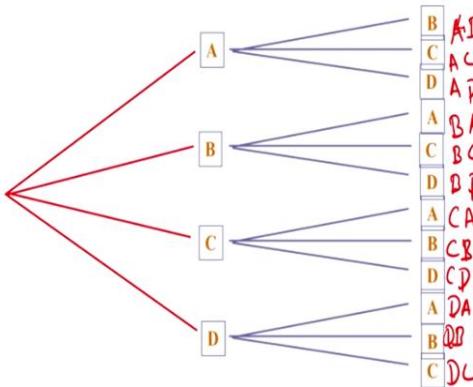
15

See this listing experiment randomly select without replacement 2 families from the residents of the town, so each order the pair in the sample spaces elementary event for example D, C. So

what are the different possibility look at this table A B, A C, A D, B A, B C, B D, C A, C B, C D so, these are the listing the sample space what is that we have to select 2 families from the residents.

(Refer Slide Time: 08:58)

Sample Space: Tree Diagram for Random Sample of Two Families



16

So, here we will do without replacement, without replacement means suppose A it owned by again A, if it is B it owned by again B, once A is taken we are not selecting another A, so without replacement the same thing the another way to express the sample spaces with the help of tree diagram, it is a tree diagram is very useful and easy to understand. For example A B C D there are 4 families we can have combination A B we can combination A C, A D, B A, B C, B D, C A, C B, C D, D A, D B, D C. So, this is the easy which is the different sample space because tree diagram is easy to understand.

(Refer Slide Time: 09:55)

Sample Space: Set Notation for Random Sample of Two Families

- $S = \{(x,y) \mid x \text{ is the family selected on the first draw, and } y \text{ is the family selected on the second draw}\}$
- Concise description of large sample spaces



17

Now the set notation for random sample of 2 families so $S = \{(X, Y), X \text{ is the family selected on the first draw, and } Y \text{ is the family selected on the second draw}\}$. It is the concise description of larger sample spaces in mathematics they use this kind of notations.

(Refer Slide Time: 10:15)

Sample Space

- Useful for discussion of general principles and concepts

Listing of Sample Space

(A,B), (A,C), (A,D),
(B,A), (B,C), (B,D),
(C,A), (C,B), (C,D),
(D,A), (D,B), (D,C)

Venn Diagram



18

You see the sample space can be shown in terms of Venn diagram, so this is a list of sample space see that this is a different dot express different sample space

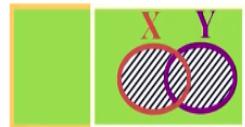
(Refer Slide Time: 10:26)

Union of Sets

- The union of two sets contains an instance of each element of the two sets.

$$X = \{1,4,7,9\}$$
$$Y = \{2,3,4,5,6\}$$
$$X \cup Y = \{1,2,3,4,5,6,7,9\}$$

$$C = \{IBM, DEC, Apple\}$$
$$F = \{Apple, Grape, Lime\}$$
$$C \cup F = \{IBM, DEC, Apple, Grape, Lime\}$$



19

Then we will go to the another concept union of sets, the union of 2 sets contains an instance of each element of the 2 sets for example X is 1,4,7,9 one set, Y is another set 2,3,4,5,6 So, if you want to know X union Y just we have to combine 1,2,3,4,5,6,7,9 similarly we look at the Venn diagram X is 1 Y is 1 if you want to know union combining both events and other examples say C IBM, DEC, Apple that is the C set there is the another set F Apple, Grape, Lime suppose we want to know union of set C and F so we are to take IBM, DEC, Apple, Apple is coming in both sets we are taking only one grape lime, this is the union.

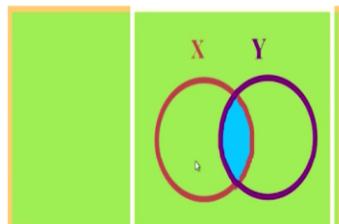
(Refer Slide Time: 10:15)

Intersection of Sets

- The intersection of two sets contains only those element common to the

$$X = \{1,4,7,9\}$$
$$Y = \{2,3,4,5,6\}$$
$$X \cap Y = \{4\}$$

$$C = \{IBM, DEC, Apple\}$$
$$F = \{Apple, Grape, Lime\}$$
$$C \cap F = \{Apple\}$$



20

We go for intersection suppose for $X = 1, 4, 7, 9$ $Y = 2, 3, 4, 5, 6$ if you want to know common, intersect is 2 sets contain only those elements common so here the forest common in X and Y so

X intersection Y is 4, for example, C and F, in C we have IBM, DEC, APPLE F is APPLE, GRAPE, LIME then C intersection F that is a common thing between set C and F is Apple, so this one, see this portion says our intersection.

(Refer Slide Time: 11:54)

Mutually Exclusive Events

- Events with no common outcomes
- Occurrence of one event precludes the occurrence of the other event

$$C = \{IBM, DEC, Apple\}$$

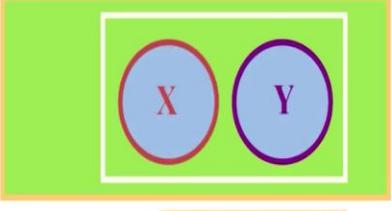
$$F = \{Grape, Lime\}$$

$$C \cap F = \{\}$$

$$X = \{1, 7, 9\}$$

$$Y = \{2, 3, 4, 5, 6\}$$

$$X \cap Y = \{\}$$



$$P(X \cap Y) = 0$$

Swayam
21

Then we will go for mutually exclusive events even with the no common outcomes is called mutually exclusive events occurrence of one event precludes the occurrence of other event for example C IBM, DEC, Apple F is Grape, Lime. So C intersection there is no common thing so that is why it is null set similarly X is 1,7,9 Y is 2,3,4,5,6 , X intersection Y there is no common set look at the Venn diagram there is no common thing so X intersection Y 0 these 2 sets are not overlapping.

So it is called mutual exclusive events. another example for this when we toss a coin there is 2 possibility to get the outcome one may be head or tail it cannot have both that is why both events are mutually exclusive event.

(Refer Slide Time: 12:46)

Independent Events

- Occurrence of one event does not affect the occurrence or nonoccurrence of the other event
- The conditional probability of X given Y is equal to the marginal probability of X.
- The conditional probability of Y given X is equal to the marginal probability of Y.

$$P(X|Y) = P(X) \text{ and } P(Y|X) = P(Y)$$



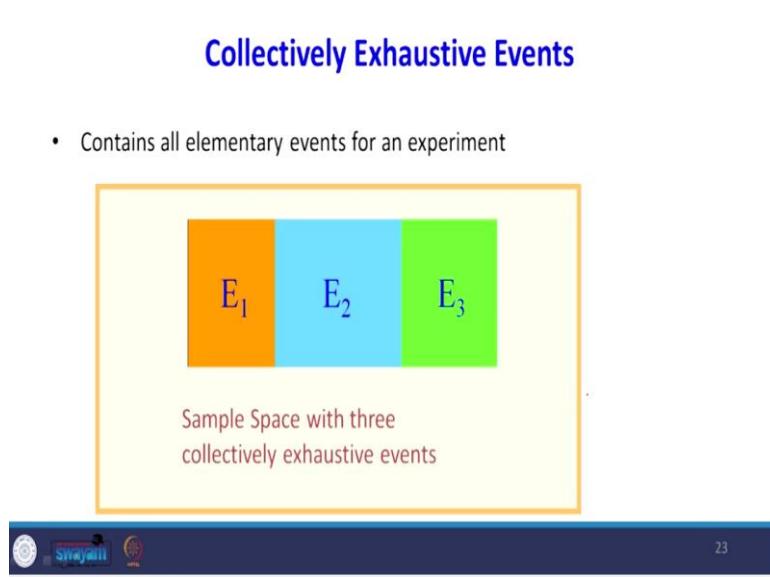
22

Then independent events, so occurrence of one event does not affect the occurrence or non occurrence of other event is called independent event, the conditional probability of X given Y is equal to the marginal probability of X the conditional probability of Y given X is equal to the marginal probability the one way we will do a small problem on this one way to test the independent event is suppose $P(X/Y) = P(X)$ and $P(Y/X) = P(Y)$, then even X and Y are called independent events will go in detail after some time but with the help of an example.

(Refer Slide Time: 13:28)

Collectively Exhaustive Events

- Contains all elementary events for an experiment



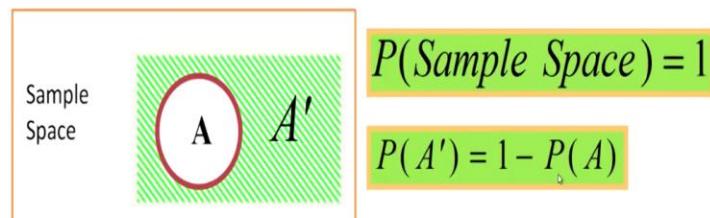
23

Collectively exhaustive event it contains all elementary events for an experiment suppose E1 E2 E3 sample space with 3 collectively exhaustive event suppose you roll your die, all possible outcome 1, 2, 3, 4, 5, 6 that is collectively exhaustive events.

(Refer Slide Time: 13:52)

Complementary Events

- All elementary events not in the event 'A' are in its complementary event.



(Refer Slide Time: 14:11)

Counting the Possibilities

- mn Rule
- Sampling from a Population with Replacement
- Combinations: Sampling from a Population without Replacement

Then complementary events and elementary events not in the A dash or is it is complimentary event you see that the $P(A)$ is there which is not there that is A dash that is called complimentary, so $P(A') = 1 - P(A)$ then counting the possibilities because in probabilities many time different combinations may come these rules may be very useful for counting different possibilities one rule is mn rule second one is sampling from a population with replacements, second one is sampling from a population without replacement.

(Refer Slide Time: 14:28)

mn Rule

- If an operation can be done m ways and a second operation can be done n ways, then there are mn ways for the two operations to occur in order.
- This rule is easily extend to k stages, with a number of ways equal to $n_1 \cdot n_2 \cdot n_3 \cdots n_k$
- Example: Toss two coins . The total umber of simple events is $2 \times 2 = 4$



26

Will go for the mn Rule if an operation can be done m ways and the second operation can be done n ways, then there are mn ways for the 2 operation to occur in order. The rule is easily can be extended to k stages, with a number of ways equal to if there are k stages n_1, n_2, n_3 there some simply we have to multiply for example toss 2 coins the total number of sample event is $2 \times 2 = 4$ because in the first coin you make a 2 possibilities second coin you make it another 2 possibilities so the total is 4 possibilities.

(Refer Slide Time: 15:07)

Sampling from a Population with Replacement

- A tray contains 1,000 individual tax returns. If 3 returns are randomly selected **with replacement** from the tray, how many possible samples are there?
- $(N)^n = (1,000)^3 = 1,000,000,000$



27

Suppose you see that another example of sampling from a population with replacement. One example is a tray contains 1000 individual tax returns if 3 returns are randomly selected with replacement from the tray, how many possible samples are there? So every time you are going

for a 3 trial; trial 1 trial 2 trial 3, in each trial there are 1000 possibilities because you can choose one from 1000, so, firstly trial 1000 second trial 1000 third trial 1000 when you multiply this is 1000 million possibilities are there with replacement.

(Refer Slide Time: 15:45)

Combinations

- A tray contains 1,000 individual tax returns. If 3 returns are randomly selected **without replacement** from the tray, how many possible samples are there?

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{1000!}{3!(1000-3)!} = 166,167,000$$



28

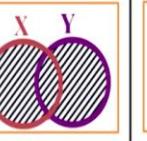
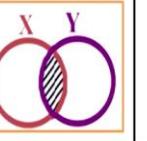
In case if you go without replacement, the same thing because without replacement what will happen the sample size will decrease. A tray contains 1000 individual tax returns in 3 returns are randomly selected without replacement from the tray. How many possible samples are there So, that is ${}^N C_n = N!/(n!(N-n)!)$

$$= 1000!/(3!(1000-3)!) = 166,167,000$$

you see the previously with replacement and going to previous light with replacement it is a 1000 million now, it is only 166 million because we are going for without replacement.

(Refer Slide Time: 16:29)

Four Types of Probability

Marginal	Union	Joint	Conditional
$P(X)$ The probability of X occurring 	$P(X \cup Y)$ The probability of X or Y occurring 	$P(X \cap Y)$ The probability of X and Y occurring 	$P(X Y)$ The probability of X occurring given that Y has occurred 



29

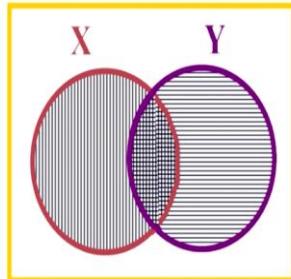
There are different types of probability say we can call it as a marginal probability union probability joint probability and conditional probability. Then what is the rotation model probabilities simple one probability P of X , so, how it is expressed in terms of Venn diagram, see this one, so marginal probability the union probability is the X union Y , the probability of X or Y counting's, the joint probability or common probability, the probability of X and Y occurring together the middle portions.

Then conditional probability, the probability of X occurring given that Y has occurred here there are 2 events, the probability of the outcome of X is depending upon the outcome of Y . So we have to read the probability of X given that Y has occurred. So this is the Venn diagram notation for expressing the conditional probability

(Refer Slide Time: 17:23)

General Law of Addition

$$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$$



Then we will go for general law of addition so, $P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$

(Refer Slide Time: 17:35)

Design for improving productivity?



Will take a small example, from that example, will understand the concept of probabilities a company is going for improving the productivity of the particular unit. They are coming with a new design one design is layout design for example, layout design, one design will reduce the noise, that is one option that is second design that will give you more storage space, so we are going to ask from the employees what kind out of these 2 designs which design will improve the productivity.

(Refer Slide Time: 18:13)

Problem

- A company conducted a survey for the American Society of Interior Designers in which workers were asked which changes in office design would increase productivity.
- Respondents were allowed to answer more than one type of design change.

Reducing noise would increase productivity	70 %
More storage space would increase productivity	67 %

You see the problem, a company conducted a survey for the American Society of interior design in which workers were asked which changes in the office design would increase productivity, there are 2 design is there one is the one design will reduce the noise another design will improve the storage space. The responders were allowed to answer more than one type of design changes.

So this table shows the outcome so 70% as the people have responded that reducing noise would increase the productivity 67 percentage of the respondents responded that more storage space would increase the productivity there is the 2 design So, we are asking there from the respond which design will improve the productivity.

(Refer Slide Time: 19:04)

Problem

- If one of the survey respondents was randomly selected and asked what office design changes would increase worker productivity,
 - what is the probability that this person would select reducing noise or more storage space?

Swayam	33
--------	----

Suppose, if one of the survey respondents were randomly selected and asked what office design change would increase workers productivity, otherwise, what is the probability that this person would select reducing noise or it designed which is helpful for providing more storage space out of this that reduce, out of 2 options.

(Refer Slide Time: 19:29)

Solution

- Let N represent the event "reducing noise."
- Let S represent the event "more storage/ filing space."
- The probability of a person responding with N or S can be symbolized statistically as a union probability by using the law of addition.

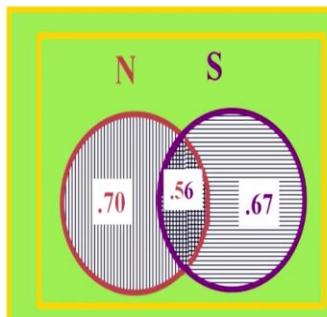
$$P(N \cup S)$$

So, let N represent is the event reducing noise that means you are choosing that design S represents the event more storage space, yes, that is the another options. The probability of person responding to N or S can be symbolized statistically as a union probability by using law of addition that is a $P(N \cup S)$.

(Refer Slide Time: 19:56)

General Law of Addition -- Example

$$P(N \cup S) = P(N) + P(S) - P(N \cap S)$$



$P(N) = .70$
 $P(S) = .67$
 $P(N \cap S) = .56$

$$P(N \cup S) = .70 + .67 - .56$$

$$= 0.81$$

So, see that P of N union S is because we know they asked for the our the formula $P(N \cup S) = P(N) + P(S) - P(N \cap S)$. So, the P of N is 70% P of S is 67% those who have told S for both the designs is 0.56. So, when you substitute these values in the formula, so, we are getting 0.81 that is 81% of the people have told both the designs will increase the productivity

(Refer Slide Time: 20:33)

		Increase Storage Space		Total
Noise Reduction	Yes	No		
	Yes	.56	.14	.70
	No	.11	.19	.30
Total		.67	.33	1.00

36

What you have solved with the help of Venn diagram in previous in the previous slide can be solved with the help of contingency table. This contingency table is so helpful just we have to make a table. For example, in rows I have taken noise reduction, the noise reduction say 70% people have told yes, so, remaining 30 might have told no. 70 30 in the column I have taken increasing storage space design in the 67% total they have told that increasing storage space would increase the productivity.

So remaining 33 might have told no. So, the 0.56 is intersection people have told both yes for both the design that is for noise reduction and increasing storage space. So, once if you know this 0.56 the remaining things you can simply you can subtract it $0.70 - 0.56 = 0.14$ that is those who have told no to storage space and yes to noise reduction then if you subtract from $0.67 - 0.56$, will get 0.11, $0.33 - 0.14$ we will get 0.19 so from this table we can read a lot of information's.

(Refer Slide Time: 21:49)

Joint Probability Using a Contingency Table

Event	Event		Total
	B ₁	B ₂	
A ₁	P(A ₁ and B ₁)	P(A ₁ and B ₂)	P(A ₁)
A ₂	P(A ₂ and B ₁)	P(A ₂ and B ₂)	P(A ₂)
Total	P(B ₁)	P(B ₂)	1

Joint Probabilities

Marginal (Simple) Probabilities

37

Whatever it has no this is for example, event A A1 A2 event B1 B2 this in the rows, you see that? It is in the columns whatever cell inside the cell this portion is called the joint probabilities P (A1 ∩ B1) whatever it is the extreme side of the table see that this called marginal probabilities, this is a notation that traditionally they follow whatever the inside the cell it is a combination of both event that is the: it is the joint probabilities, whatever extremely and it is a marginal probability extreme side of the table.

(Refer Slide Time: 22:26)

Office Design Problem - Probability Matrix

		Increase Storage Space		Total
		Yes	No	
Noise Reduction	Yes	.56	.14	.70
	No	.11	.19	.30
Total		.67	.33	1.00

$$\begin{aligned}
 P(N \cup S) &= P(N) + P(S) - P(N \cap S) \\
 &= .70 + .67 - .56 \\
 &= .81
 \end{aligned}$$

38

The same thing, suppose if you want to know the same answer with the help of this contingency table, we want to know how much percentage of the people who have agreed for both the design that is N and S. So, P (N) + P (S) - P (N ∩ S). So this value directly we can read it from the table.

So $P(N)$ is 0.70 + $P(S)$ is 0.67 - $P(N \text{ intersection } S)$ is 0.56, so when you do that we getting 0.81.

(Refer Slide Time: 22:57)

Law of Conditional Probability

$$P(N) = .70$$

$$P(N \cap S) = .56$$

$$P(S|N) = \frac{P(N \cap S)}{P(N)}$$

$$= \frac{.56}{.70}$$

$$= .80$$

Swayam

Then we will go for conditional probability the probability of N is 0.70 those who are good both the design engineers is .56 suppose if you want to know $P(S \setminus N) = P(N \text{ intersection } S) / P(N)$ so this value I will explain this conditional probability in detail later, but now you take this one, so, P of N intersection S from the previous table you can find out the 0.56. You can look at the Venn diagram also, P of N is 0.7 and you substitute getting 0.8

(Refer Slide Time: 23:37)

Office Design Problem

		Increase Storage Space		
		Yes	No	Total
Noise Reduction	Yes	.56	.14	.70
	No	.11	.19	.30
Total	.67	.33	1.00	

$$P(\bar{N} | S) = \frac{P(\bar{N} \cap S)}{P(S)} = \frac{.11}{.67} = .164$$

Swayam

The same office design problem you see that there is another conditional probability $P(N/S)$ that means, those who are told No to noise reduction, but they are agreed for storage design. So,

this is the conditional probability. So, we have to multiply $P\left(\frac{\bar{N}}{S}\right) = \frac{P(\bar{N} \cap S)}{P(S)}$ is this point because

$N \cap S = 0.11$ divided by $P(S) = 0.67$ that = 0.164.

(Refer Slide Time: 24:14)

Problem

- A company data reveal that 155 employees worked one of four types of positions.
- Shown here again is the raw values matrix (also called a contingency table) with the frequency counts for each category and for subtotals and totals containing a breakdown of these employees by type of position and by sex.



41

We will take another small problem will explain the concept of probabilities with the help of the problem. A company data revealed that 155 employees worked on 4 types of positions the table is shown in the next slide raw value of matrix also called the contingency table with the frequency counts of each category and sub totals and totals containing breakdown of these employees by type of position and by sex.

(Refer Slide Time: 24:46)

Contingency Table

COMPANY HUMAN RESOURCE DATA

		Sex		11	
		Male			
Type of Position	Managerial	8	3		
	Professional	31	13	44	
	Technical	52	17	69	
	Clerical	9	22	31	
		100	55	155	



42

You see that look at this table in rows, the type of position they hold in their organization, whether they are working in managerial position, professional or technical or clerical in the column sex whether male or female, you see the intersection of managerial and male 8 that represents both that is more managerial working in a managerial position the same time they are male. So, that is our joint values here the only count joint counts the extreme right or the bottom of the table we are given the total counts.

(Refer Slide Time: 25:23)

Solution

- If an employee of the company is selected randomly, what is the probability that the employee is female or a professional worker?

$$P(F \cup P) = P(F) + P(P) - P(F \cap P)$$

$$P(F \cup P) = .355 + .284 - .084 = .555.$$



43

Now, if an employee of the company is selected randomly, what is the probability that the employee is female or professional worker, so, what we have to do? So, we are going to find out $P(F) + P(P) - P(F \cap P)$. So, $P(F)$ is when you go to the previous one $P(F)$ is so, when you

$55/155 = 0.355$, $P(P)$ is when $44/155 = 0.824$, minus $P(F \cap P)$. that is F intersection $P = 13/155$ you will get 0.084

= 0.55.

There is another problem.

(Refer Slide Time: 26:15)

Problem

- Shown here are the raw values matrix and corresponding probability matrix for the results of a national survey of 200 executives who were asked to identify the geographic locale of their company and their company's industry type.
- The executives were only allowed to select one locale and one industry type.



Shown here are the raw value matrix and corresponding probability matrix of the result of a national survey of 200 executives who were asked to identify geographic location of their company and their company's industry type. The executives were only allowed to select one location and one industry type, because it is not possible same person working different location different industries, because one person can work only one type of industry will conclude that in this session.

We have seen different types of probability how to assign probability and different counting rules say mn rules with replacement, without replacement. And different terms, which you are frequently we are going to use in this course that is the event, joint probability, marginal probability and so on, then you have taken one sample problem with the help of sample problem, we have seen how to find out union of 2 events that is a joint probability $P(A \cup B)$, then intersection $P(A \cap B)$.

Then how to find out the marginal probability, then how to find out the conditional probability, but that will close and continue with the next lecture. Thank you very much

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 07
Introduction to Probability - II

Dear Students, we will continue that the concept of probability in lecture number 7 will take one example then we will try to understand the concept of marginal probability, joint probability, and conditional probability in this problem. The problem is a company data reveal that 155 employees worked in a 4 types of positions. Shown here again is the raw values matrix also called the contingency table with the frequency count for each category.

And subtotals and totals containing a breakdown of these employees by type of position and the sex. Look at this contingency table in the row it is given what kind of position they are holding whether managerial professional technical clerical,

(Refer Slide Time: 01:20)

Solution

- If an employee of the company is selected randomly, what is the probability that the employee is female or a professional worker?

$$P(F \cup P) = P(F) + P(P) - P(F \cap P)$$

$$P(F \cup P) = .355 + .284 - .084 = .555.$$

In column what is their sex, suppose when an employee of the company is selected randomly, what is the probability that the employee is female, in the contingency table row represents the type of position, column represents sex, the type of position they may hold whether they can have the they can work as a managerial position, professional position, technical clerical, we

asked their sex also in the datasets, the question is if an employee of the company is selected randomly.

What is the probability that the employee is female or professional worker that is what is the $P(F \cup P)$? F is the female, P is the professional. So, as per the law of addition of probability

$$P(F \cup P) = P(F) + P(P) - P(F \cap P),$$

$P(F)$ we can find out and going to the previous slide.

(Refer Slide Time: 02:22)

Contingency Table

COMPANY HUMAN RESOURCE DATA

		<i>Sex</i>		155
		Male	Female	
Type of Position	Managerial	8	3	11
	Professional	31	13	44
	Technical	52	17	69
	Clerical	9	22	31
		100	55	

So, the $P(F)$ is there are 55 females total there are 155 So, when $= 55/155$, you will get 0.355 then $P(P)$, going to previous slide $P(P)$, is probability of professionals, there are 44 professionals, $44/155$ that will give you 0.284 then, minus $P(F \cap P)$, that is for female related same time their working in a professional type of position that is 13. So, $13/155$ will give you 0.084

$$P(F \cup P) = 0.555.$$

(Refer Slide Time: 03:10)

Problem

- Shown here are the raw values matrix and corresponding probability matrix for the results of a national survey of 200 executives who were asked to identify the geographic locale of their company and their company's industry type.
- The executives were only allowed to select one locale and one industry type.

Shown here are the raw value matrix and corresponding probability matrix of the result of a national survey of 200 executives who are asked to identify the geographical location of their company and their company's industry type. So, there asking 2 question what is their company's geographic location and what kind of industry they are working. The executives were only allowed to select one location and one industry because they can work only one industry in one location.

(Refer Slide Time: 03:45)

		RAW VALUES MATRIX				
		Geographic Location				
		Northeast	Southeast	Midwest	West	
Industry Type	Finance A	D	24	10	8	14
	Manufacturing B	E	30	6	22	12
	Communications C	F	28	18	12	16
		82	34	42	42	200

This table shows 0 there is a industry type maybe finance manufacturing communications, calling it ABC. The geographic location may be Northeast, Southeast, Midwest and West, So for example, in the finance A, there are 56 people are working manufacturing B there are 70 people

are working, in the communications 74 in the Northeast location, there are 82 people, in Southeast location 34 people and Midwest location 42 and West 42.

(Refer Slide Time: 04:21)

Questions

- a. What is the probability that the respondent is from the Midwest (F)?
- b. What is the probability that the respondent is from the communications industry (C) or from the Northeast (D)?
- c. What is the probability that the respondent is from the Southeast (E) or from the finance industry (A)?

The question is, what is the probability that the respondent is from the Midwest? Directly we can read this answer from the table, second question is what is the probability that the respondent is from the communications industry or from the northeast? So here the addition of the 2 probability that is a $P(C)$ and $P(D)$. The third question is what is the probability that the respondent is from the Southeast or from the finance industry?

(Refer Slide Time: 04:55)

		Geographic Location				
		Northeast	Southeast	Midwest	West	
Industry Type	Finance A	.12	.05	.04	.07	.28
	Manufacturing B	.15	.03	.11	.06	.35
	Communications C	.14	.09	.06	.08	.37
		.41	.17	.21	.21	1.00

So from the given table, we find out the probability that means we have divided each element in the cell divided by the gross total. After dividing that we got 0.12, 0.05, 0.04, 0.04 Now this is a matrix conditional probability. From this table, we can pick up whatever answer we wanted to get for answering that question

(Refer Slide Time: 05:20)

Mutually Exclusive Events

Type of Position	Gender		Total
	Male	Female	
Managerial	8	3	11
Professional	31	13	44
Technical	52	17	69
Clerical	9	22	31
Total	100	55	155

$$\begin{aligned}
 P(T \cup C) &= P(T) + P(C) \\
 &= \frac{69}{155} + \frac{31}{155} \\
 &= .645
 \end{aligned}$$

For example. I am going back, suppose what is the probability that the respondent is from the Midwest? So I am going to next slide Midwest is F. So, it is a 0.21 going back, what is the probability that the respondent is from the communication industry or Northeast? So you have to see $P(C \cup D) = P(C) + P(D) - P(C \cap D)$ you can find out then what is the probability that respondent is from southeast or the same thing? $P(E \cup A) = P(E) + P(A) - P(E \cap A)$ These values you can directly pick up from the previous table from, this table now we will go for mutually exclusive event suppose okay suppose if you want to find out $P(T \cup C)$ here T is the those who are technical professional P of T union C is $P(T) + P(C)$ because generally it will be minus $P(T \cap C)$, but that is not possible, because a person cannot work in 2 industry at a time.

So, it is mutually exclusive event in the mutually exclusive event in the intersection will become 0 that is a $P(T \cup C) = P(T) + P(C)$ that another term that is minus $P(T \cap C)$ will become 0. So, $P(T)$ is a $69/155$, $+ P(C)$ is $31/155$ when you simplify it is 0.645. This is example of mutually exclusive event.

(Refer Slide Time: 07:00)

Mutually Exclusive Events

Type of Position	Gender		Total
	Male	Female	
Managerial	8	3	11
Professional	31	13	44
Technical	52	17	69
Clerical	9	22	31
Total	100	55	155

$$\begin{aligned}P(P \cup C) &= P(P) + P(C) \\&= \frac{44}{155} + \frac{31}{155} \\&= .484\end{aligned}$$

There is another $P(P \cup C) = P(P) + P(C)$ there would not be any intersections even that formula which I told the 2 slides before also that the intersection component will become 0 because the person cannot work in 2 industries, there is an example of mutually exclusive event.

(Refer Slide Time: 07:18)

Law of Multiplication

$$P(X \cap Y) = P(X) \cdot P(Y|X) = P(Y) \cdot P(X|Y)$$

In the law of multiplication $P(X \cap Y) = P(X) \cdot P(Y|X) = P(Y) \cdot P(X|Y)$ is a law of multiplication. What will happen if event X and Y are independent event? Simply we can multiply $P(X)$ into $P(Y)$, that you will see later.

(Refer Slide Time: 07:42)

Problem

- A company has 140 employees, of which 30 are supervisors.
- Eighty of the employees are married, and 20% of the married employees are supervisors.
- If a company employee is randomly selected, what is the probability that the employee is married and is a supervisor?

You will see another problem. A company has a 140 employees of which 30 are supervisors 80 of the employees are married, and 20% of the married employees are supervisors. If a company employee is randomly selected, what is the probability that the employee is married and is a supervisor? Wherever this kind of problem comes, if you are able to construct the contingency table whatever question being asked you can pick up from there directly. So, from the given data.

(Refer Slide Time: 08:15)

		Married		
		Y	N	Sub total
Supervisor	Y	0.1143		30
	N			110
	Sub total	80	60	140

First we will construct a contingency table in the contingency table you see there are 140 employees out of which 30 are supervisors out of 140, 80 people are married and for example, those who are married at the same time supervisors, what do you do that is you have to multiply that how will you that if you multiply 80 into 0.2 divided by 140 you will get this answer.

(Refer Slide Time: 08:43)

$$P(M) = \frac{80}{140} = 0.5714$$

$$P(S|M) = 0.20$$

$$\begin{aligned} P(M \cap S) &= P(M) \cdot P(S|M) \\ &= (0.5714)(0.20) = 0.1143 \end{aligned}$$

So, for example, you see that how we are getting that we are the previously they will get 0.1143 we will see how we are getting so, we know the probability of married people 80 /140. This is given those who are supervisors at the same time married. 0.2 that is a 20% is given. If you want to know those who are married at the same time they are supervisors. So, we have to use conditional probability $P(M) \times P(S|M)$. So, $P(S|M)$ is known, 0.20 is given. So $P(M \cap S) = 0.1143$.

(Refer Slide Time: 09:32)

Law of Multiplication

		Married	
		Yes	No
Supervisor	Yes	.1143	.8857
	No	.4571	.54286
Total	.5714	.4286	1.00

$$\begin{aligned} P(S) &= 1 - P(\bar{S}) \\ &= 1 - 0.2143 = 0.7857 \\ P(\bar{M} \cap \bar{S}) &= P(\bar{S}) - P(M \cap \bar{S}) \\ &= 0.7857 - 0.4571 = 0.3286 \\ P(M \cap S) &= P(M) - P(M \cap \bar{S}) \\ &= 0.5714 - 0.1143 = 0.4571 \\ P(\bar{M} \cap S) &= P(S) - P(M \cap S) \\ &= 0.2143 - 0.1143 = 0.1000 \\ P(\bar{M}) &= 1 - P(M) \\ &= 1 - 0.5714 = 0.4286 \end{aligned}$$

You see that once you know that one cell in the contingency table filling the remaining cell is so easy. And whatever value you want to pick up we can pick up for example, I have filled the first

0.1143 I know what is the row total and column total from that I can subtract it I can get the remaining rows that is a application of contingency table. Suppose $P(S) = 1 - P(S_{\bar{}})$, that we know that $P(S_{\bar{}})$, that is 0.7857.

I am saying this one, this location, this location 0.7857 if you want to know $P(M_{\bar{}} \cap S_{\bar{}})$ that means, those who are not M those are not married. At the same time, who are not supervisors, so, that is nothing but this location 0.326 those are not married the same time they are not supervisor this locations, $P(M_{\bar{}} \cap S_{\bar{}})$ those are married, but not the supervisors.

So, that is a $P(M) - P(M \cap S)$. So, this value is 0.5714 minus this one. So, we will get this point, nothing but in the contingency table if you know one cell the remaining this can be found out

(Refer Slide Time: 10:54)

Special Law of Multiplication for Independent Events

- General Law

$$P(X \cap Y) = P(X) \cdot P(Y | X) = P(Y) \cdot P(X | Y)$$

- Special Law

If events X and Y are independent,
 $P(X) = P(X | Y)$, and $P(Y) = P(Y | X)$.
Consequently,
 $P(X \cap Y) = P(X) \cdot P(Y)$

The special law multiplication for independent events general law is if $P(X \cap Y) = P(X) \cdot P(Y | X) = P(Y) \cdot P(X | Y)$. special law, that is if X and Y are independent. So, $P(X) = P(X | Y)$, because, when the event X and Y are independent, the outcome of X is not depending on outcome of Y. So, $P(X | Y)$ will become $P(X)$ itself. So, similarly, $P(Y)$ and $P(X | Y)$ if not independent $P(Y | X)$ will become $P(Y)$ itself and you substitute the there. So, $P(X \cap Y)$ will be $P(X) \cdot P(Y)$, only when the both even are independent.

(Refer Slide Time: 11:45)

Law of Conditional Probability

- The conditional probability of X given Y is the joint probability of X and Y divided by the marginal probability of Y.

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{P(Y|X) \cdot P(X)}{P(Y)}$$

This also law of conditional probability, this also we have seen previously also the conditional probability of X given Y is joint probability of X and Y divided by the marginal probability of Y So joint properties intersection the $P(X \text{ intersection } Y) / P(Y)$. So this can be just by readjusting $(P(Y / X) \times P(X)) / P(Y)$.

(Refer Slide Time: 12:14)

Conditional Probability

- A conditional probability is the probability of one event, given that another event has occurred:

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)} \quad \xrightarrow{\text{The conditional probability of A given that B has occurred}}$$

$$P(B | A) = \frac{P(A \text{ and } B)}{P(A)} \quad \xrightarrow{\text{The conditional probability of B given that A has occurred}}$$

Where $P(A \text{ and } B)$ = joint probability of A and B

$P(A)$ = marginal probability of A

$P(B)$ = marginal probability of B

Little detailed explanation on conditional probability, A conditional probability is the probability of one event, given that another event has occurred suppose if I say $P(A | B)$. So, first you have to find the intersection of $P(A \cap B)$ then divide by $P(B)$. So the conditional probability of A given that B has occurred. This is an explanation for this. Suppose, if you want to know $P(B)$ given A has occurred, so $P(A \cap B) / P(A)$

where P of A and B equal to join probability of A and B. So P (A) is marginal property of event A, P (B) is marginal property event B

(Refer Slide Time: 12:54)

Computing Conditional Probability

- Of the cars on a used car lot, 70% have air conditioning (AC) and 40% have a CD player (CD). 20% of the cars have both.
- What is the probability that a car has a CD player, given that it has AC?
- We want to find $P(CD | AC)$.

We will take an example how to find out the conditional probability of the cars on a used car lot, 70% have air conditioning Air Conditioning and 40% have CD player. 20% of the cars have both. So, what is the probability that a car has a CD player, given that it has AC that means, we want to find out P of CD given that AC is there.

(Refer Slide Time: 13:19)

Computing Conditional Probability

	CD	No CD	Total
AC	0.2	0.5	0.7
No AC	0.2	0.1	0.3
Total	0.4	0.6	1.0

$$P(CD | AC) = \frac{P(CD \text{ and } AC)}{P(AC)} = \frac{0.2}{0.7} = 0.2857$$

Given AC, we only consider the top row (70% of the cars). Of these, 20% have a CD player. 20% of 70% is about 28.57%.

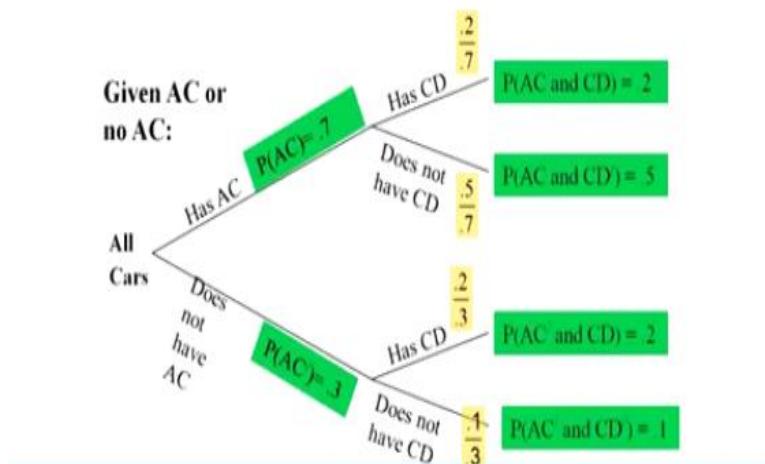
As I told you just to draw the contingency table because all the values are given? So, what is the value I am going back see for example 70% of the cars having AC So, this value I am going

back 40% of the cars having CD player So, this value and you subtract minus 1 will get that one another information is given 20% of the car have both like by see that 0.2 this value. So, once you know these values other cells can be find out.

So, if you want to know $P(CD|AC)$ as per the definition, P of CD and AC divided by $P(AC)$ so, this is a 0.2 this is by $P(AC)$ is by 0.7. So, this is 0.2857. So, given the AC we only consider the top row 70% of the cars of these and 20% CD player, so 20% of % is 28.57% okay there so, we are getting the conditional probability.

(Refer Slide Time: 14:25)

Computing Conditional Probability: Decision Trees



So, this conditional probability can be explained with the help of a tree diagram, because the tree diagram is easy to visualize. So, having AC, having not AC, having CD, having not CD having CD having not CD. So, 0.7 0.2 0.5 0.2 0.1 so, if you want to know having CD, so you have to divide 0.2 divided by 0.7. For example, if you want to know this this arc, so, this is a 0.5 divided by 0.7 and so on because a tree diagram is very easy to understand.

(Refer Slide Time: 15:12)

Independent Events

- If X and Y are independent events, the occurrence of Y does not affect the probability of X occurring.
- If X and Y are independent events, the occurrence of X does not affect the probability of Y occurring.

If X and Y are independent events

$$P(X|Y) = P(X), \text{ and}$$

$$P(Y|X) = P(Y).$$

Then we will see the definition of independent event, if an X and Y are independent events, the occurrence of Y does not affect the probability of X occurring, so, X and Y are not connected. Similarly, if X and Y are independent events, the occurrence of Y does not affect the probability of X occurring, you see that P of if X and Y are independent $P(X \setminus Y) = P(X)$, $P(Y \setminus X) = P(Y)$ this we have seen the previous also.

(Refer Slide Time: 15:44)

Statistical Independence

- Two events are **independent** if and only if:

$$P(A | B) = P(A)$$

- Events A and B are independent when the probability of one event is not affected by the other event

This is another example 2 events are independent This is the condition $P(A \setminus B) = P(A)$. So, this condition is for testing independent, even A and B are independent. When the probability of one event is not affected by other event.

(Refer Slide Time: 16:05)

Independent Events Demonstration

		Geographic Location			
		Northeast D	Southeast E	Midwest F	West G
Industry Type	Region				
		D	E	F	G
Finance A	D	.12	.05	.04	.07
Manufacturing B	E	.15	.03	.11	.06
Communications C	F	.14	.09	.06	.08
	G	.41	.17	.21	.21
					1.00

Test the matrix for the 200 executive responses to determine whether industry type is independent of geographic location.

So, we will take one example will check the practical application of this concept of independent events. This also this data previously given. So, we have asked what kind of industry are working, whether you will finance manufacturing communications that we asked to the geographical locations. Now, you see that the question is tested the matrix for the 200 executive respondents to determine whether the industry type is independent of geographical location that means, we were to find out is there any dependency between the geographical location and what kind of industry.

For example, in India, if you look at there, most of you know, software companies are in south. So, is there any connection between the geographical location and kind of industry which are located. We will take this Example finance and the best region, so, when you go this,

(Refer Slide Time: 17:05)

Independent Events Demonstration Contd...

$$P(A|G) = \frac{P(A \cap G)}{P(G)} = \frac{0.07}{0.21} = 0.33 \quad P(A) = 0.28$$
$$P(A|G) = 0.33 \neq P(A) = 0.28$$

If you want to know $P(A|G) = P(A \cap G)/P(G)$, $P(A \cap G)$ we can directly read from the table 0.07 this one this value then $P(G) 0.21$ directly we can read from the table. So, what you do that value is 0.33, but $P(A)$ when you look at $P(A)$, so, $P(A)$ is 0.28. So, now what is happening, the $P(A|G)$ is not equal to $P(A)$. If it is equal, both are independent, since it is not equal, there is a kind of dependency between the geographical location and the type of industry which are located there. So, this is a one way to test the independency.

(Refer Slide Time: 18:06)

Independent Events

	D	E	
A	8	12	20
B	20	30	50
C	6	9	15
	34	51	85

$$P(A|D) = \frac{8}{34} = 0.2353$$
$$P(A) = \frac{20}{85} = 0.2353$$
$$P(A|D) = P(A) = 0.2353$$

For example, you take for another example, because A given D, this is another example. So, $P(A \cap D)$ is 8 here the actual count is given. Any way you can you can do both the way also $P(A \cap D)$ is 8 divided by $P(D)$, $P(D)$ is 34. So, we are getting this value, but you see the $P(A)$

is 20 divided by 85, 20 divided by 85. So, both $P(A \setminus D)$ and $P(A)$ are same. So, these are independent events. Then example, if at the same $P(A \setminus D)$ equal to $P(A)$ both events are independent this is the way to test the independent.

(Refer Slide Time: 18:55)

Revision of Probabilities: Bayes' Rule

- An extension to the conditional law of probabilities
- Enables revision of original probabilities with new information

$$P(X_i|Y) = \frac{P(Y|X_i)P(X_i)}{P(Y|X_1)P(X_1) + P(Y|X_2)P(X_2) + \dots + P(Y|X_n)P(X_n)}$$

Next we are going to an important application that is a Bayes ruler Bayes theorem, it is used to revise the probabilities, it has lot of applications in higher level of probability theory. And extension to the conditional law of probabilities enables revision of original probability with the new information's. So, $P(X \setminus Y)$ equal to $P(Y \setminus X) \times P(X)$ divided by the summation of this one I will tell you the net the next slides

(Refer Slide Time: 19:29)

$$P(X/Y) = \frac{P(X \cap Y)}{P(Y)}$$

$$P(Y/X) = \frac{P(X \cap Y)}{P(X)}$$

$$P(X/Y) P(Y) = P(Y/X) \cdot P(X)$$

$$P(Y/X) = \frac{P(X/Y) \cdot P(Y)}{P(X)}$$

For example, supposed see that $P(X|Y)$ So, this can be written as $P(X \cap Y)/P(Y)$ this can be written as P of x intersection y divided by P of x you look at this here also P of x intersection y this also be x intersection y . So, this can be written as $P(x|y)$ multiplied by $P(y)$ equal to $P(y|x)$ multiplied by $P(x)$, you see that if I look at this suppose, I know $P(x|y)$.

Suppose, if i want to know the reverse of this that is P of y by x , you see that I know P of x by y , I am getting reverse of that that is P of y by x . So, from this you can write it P of y by x is nothing but P of x by y multiplied by P of y divided by P of x . Here the P of x is only because here only 2 outcome there are if there are more outcomes here the sigma of $P(x)$ will come, the sigma of $P(x)$ is nothing but different combination of joint probabilities that we will see with the help of an example

(Refer Slide Time: 21:12)

Problem

- Machines A, B, and C all produce the same two parts, X and Y. Of all the parts produced, machine A produces 60%, machine B produces 30%, and machine C produces 10%. In addition
 - 40% of the parts made by machine A are part X.
 - 50% of the parts made by machine B are part X.
 - 70% of the parts made by machine C are part X.
- A part produced by this company is randomly sampled and is determined to be an X part.
- With the knowledge that it is an X part, revise the probabilities that the part came from machine A, B, or C.

This is a very typical example machine A and B, and C all produce the same 2 parts X and Y. Of all of the parts produced, machine A produces 60% machine B produces 30%, and machine C produces 10%. In addition 40% of the parts made by machine A are part X 50% of the parts made by machine B are part X 70% of the parts made by machine C is part X. A party produced by this company is randomly sampled and determined to be an X part with the knowledge that it is an X part. revise the probabilities that the part came from machine A, B, and C First, we will solve this with the help of a tabular format.

(Refer Slide Time: 22:01)

Event	Prior $P(E_i)$	Conditional $P(X E_i)$	Joint $P(X \cap E_i)$	Posterior
A	.60	.40	$(.60)(.40) = .24$	$\frac{.24}{.46} = .52$
B	.30	.50	$.15$	$\frac{.15}{.46} = .33$
C	.10	.70	$\frac{.07}{P(X) = .46}$	$\frac{.07}{.46} = .15$

For example, there are 3 mission is there mission A and BC that 60% of that part was produced by machine A, 30% was produced by machine B, 10% is by C previously we have seen how that formula for conditional probability has come now, I will tell you an application of Bayes theorem.

(Refer Slide Time: 22:29)

$$\begin{array}{cc}
 \text{A} & \text{B} \\
 40\% & 60\% \\
 2\% & 3\%
 \end{array}$$

Suppose there are 2 say there are 2 supplier, supplier A supplier B, I know that, say the 40% of the product supplied by supplier A, remaining 60% supplied by supplier B from my past experience. I know that the 2% out of 40% is 2% will be defective product which are supplied by supplier A from supplier B, I know from my past experience he used to supply 3% of defective products out of 60.

By using their products that I have assembled a new machine now the machine is not working, the machine is not working, but I want to know what is the probability that product was supplied by supplier A, If the machine is not working, what is the probability that the product was supplied by supplier B? So this is the application of your Bayes theorem, we will see with the help of an example.

(Refer Slide Time: 23:36)

Problem

- A particular type of printer ribbon is produced by only two companies, **Alamo Ribbon Company** and **South Jersey Products**.
- Suppose **Alamo produces 65%** of the ribbons and that **South Jersey produces 35%**.
- Eight percent of the ribbons produced by Alamo are defective and 12% of the South Jersey ribbons are defective
- A customer purchases a new ribbon. What is the probability that Alamo produced the ribbon? What is the probability that South Jersey produced the ribbon?

Different options there. The problem is a particular type of printer ribbon is produced by only 2 companies that company names are Alamo Ribbon Company and South Jersey Products. Suppose Alamo produces 60% of the ribbons and South Jersey produces 35% of the ribbons from our experience. Look at this 8% of the ribbon produced by Alamo or defective and 12% of the South Jersey Ribbons are defective from our past experience, A customer purchases a new ribbon.

What is the probability that Alamo produced the ribbon? Otherwise, what is the probability that South Jersey produced the ribbon, like in the previous example, the machine is not working, what is the probability that product was supplied by supplier A what is the probability that product was supplied by supplier B the same example.

(Refer Slide Time: 24:39)

Revision of Probabilities with Bayes' Rule: Ribbon Problem

$$\begin{aligned}
 P(\text{Alamo}) &= 0.65 \\
 P(\text{South Jersey}) &= 0.35 \\
 P(d|\text{Alamo}) &= 0.08 \\
 P(d|\text{South Jersey}) &= 0.12 \\
 P(\text{Alamo}|d) &= \frac{P(d|\text{Alamo}) \cdot P(\text{Alamo})}{P(d|\text{Alamo}) \cdot P(\text{Alamo}) + P(d|\text{South Jersey}) \cdot P(\text{South Jersey})} \\
 &= \frac{(0.08)(0.65)}{(0.08)(0.65) + (0.12)(0.35)} = 0.553 \\
 P(\text{South Jersey}|d) &= \frac{P(d|\text{South Jersey}) \cdot P(\text{South Jersey})}{P(d|\text{Alamo}) \cdot P(\text{Alamo}) + P(d|\text{South Jersey}) \cdot P(\text{South Jersey})} \\
 &= \frac{(0.12)(0.35)}{(0.08)(0.65) + (0.12)(0.35)} = 0.447
 \end{aligned}$$

Now, first you will find out the marginal probability and conditional probability. So, P of Alamo that is 65% of the product was supplied by Alamo South Jersey 35% from the past experience I know the defective parts which was supplied by Alamo supplier 0.08. Similarly, the defective products which was supplied by South Jersey is 0.12, you see that this some would not be 100, but this some will be 100 because this is a total they supply in 8% is in the 65 out of 65, 8% is defective products are supplied by Alamo person.

Now, as for the formula now, we look at this V now it is reverse Now, we know it is defective. Then we want to find out what is the probability that was supplied by Alamo. So, we look at this P (D) given by Alamo multiple by P of Alamo, look at this this component, this is the sum of all possibilities P (D) given a Alamo multiplied by P (P Alamo)+ P (D given by South Jersey) multiplied by P (South jersey) So, this 0.08 is given 0.65 is given.

So, this was combination of 0.08 and 0.65 this and this was combination of 0.35 and 0.12. So, but because we have to add these two. So, when you divide by this is 0.553 that is, if the product the ribbon is defective, then there is a 50% chance it was supplied by supplier alone. Similarly, the product is defective what is the probability that it was supplied by South Jersey, the same thing, 0.12 to multiplied by 0.35 because P of D Alamo is given then 0.08 This is the all combination this denominator same, so, 0.447 that is there is of 44.7% chance that defective product was surprised by South Jersey.

(Refer Slide Time: 27:01)

Revision of Probabilities with Bayes' Rule: Ribbon Problem

Event	Prior Probability $P(E_i)$	Conditional Probability $P(d E_i)$	Joint Probability $P(E_i \cap d)$	Revised Probability $P(E_i d)$
Alamo	0.65	0.08	0.052	$\frac{0.052}{0.094}$ = 0.553
South Jersey	0.35	0.12	<u>0.042</u> 0.094	<u>0.042</u> 0.094 = 0.447

If you look at it in the tabular form it is very easy. So, this is the first. So, this is the event Alamo South Jersey this fellow supplying 65% this is 35%, this is the conditional probability, we know that this supplier Alamo will supply 8% of defective products, this fellow will supply 12% of 2 products. So, first we have to find out the joint property 0.0 to 0.08 this one we have to add it. Then this joint property has to be divided by this 0.04. So, that will give you see that, here we know the details of P of D given the; we are finding the reverse of that P of E given by D that was the advantage of this byes theorem. Now, it is 0.094, this was 0.447.

(Refer Slide Time: 27:54)

Revision of Probabilities with Bayes' Rule: Ribbon Problem



This can be shown in the pictorial form. There is a Alamo South Jersey defective not defective. Defective 12% remaining this percentage when we multiply by 0.02 when you multiple this 0.042 when you added, we are getting 0.094. In this lecture, I have explained the example of mutually exclusive events then I have explained the multiplication, then explain the independent events then I have explained the concept of Bayes theorem, then I have explained with the help of a problem, the application of Bayes theorem with that will conclude this lecture. Thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 08
Probability Distributions

Good morning students we are entering to the 8th lecture on this course that is a data analytics with the Python. Today the topic is probability distributions. So what we are going to cover today is very interesting topic. We are going to see the some empirical distribution and its properties. Then discrete distribution in the discrete distribution we are going to see Binomial, Poisson, Hyper geometric distributions. The continuous distribution we are going to see the uniform, exponential, normal distribution.

(Refer Slide Time: 01:00)

What is a distribution?

- Describes the 'shape' of a batch of numbers
- The characteristics of a distribution can sometimes be defined using a small number of numeric descriptors called 'parameters'

First up all what is distribution? What is the purpose of studying the distribution? The distributions describe the shape of a batch of numbers that is the meaning of distribution. Suppose the different set of numbers there, you want to show what shape it follows whether it is a bell shaped, we can call it is a normal distribution. If it is forming a rectangular shape, we can call it as a uniform distribution like this that describes the shape of a batch of numbers.

The characteristics of your distribution can sometimes be defined as a small number of numerical descriptors called parameters. So each distributions characteristic is expressed with the help of its parameters. Parameter is nothing but for example normal distribution it has 2

parameter one is mean and variance with the help of that you can draw the distribution that is a parameter.

(Refer Slide Time: 01:53)

Why distribution?

- Can serve as a basis for standardized comparison of empirical distributions
- Can help us estimate confidence intervals for inferential statistics
- Form a basis for more advanced statistical methods
 - ‘fit’ between observed distributions and certain theoretical distributions is an assumption of many statistical procedures



4

Why distribution? Can serve as a basis for standardized the comparison of empirical distributions because if you want compare with phenomena with the very standard distributions we can come to know that what distribution it follows then it will help you to estimate the confidence intervals for inferential statistics that will see what is the meaning of conference interval incoming classes then form a basis for more advanced statistical methods.

For example, fit between observed distribution and certain theoretical distribution is an assumption of many statistical procedures. Suppose why we have to study the distributions, suppose we are doing your simulation for example the arrival pattern follow Poisson distributions. Suppose certain data you collected if you prove that it is arrival follow Poisson distribution already there is a mean and variance and other population parameters already defined it.

If you are a natural phenomena, you are able to compare with standard distributions that are well defined distribution parameter is there that parameter you can use as it that is a purpose of studying the distribution.

(Refer Slide Time: 03:01)

Random variable

- A variable which contains the outcomes of a chance experiment
- “Quantifying the outcomes”
- Example $X = \{1 = \text{Head}, 0 = \text{Tails}\}$
- A variable that can take on different values in the population according to some “random” mechanism
- Discrete
 - Distinct values, countable
 - Year
- Continuous
 - Mass



5

Then we will go for what is the random variable we want to construct a distribution, it is the relation between X and corresponding probability X, p of x . So here the X is nothing but random variable. A variable which contains the outcome of chance experiment is the random variable is the kind of you are quantifying the outcome suppose we task of the coin for $X = 1$ is getting head, 0 getting tails. So 1 is nothing but your random variable.

So the X , the X is the random variable, that can take the value of 1 and 0, the X value is 1 it is the head. If X value is 0 it is a tail. Variable that can take on different values in the population according to some random mechanism. So the value of 1 and 0, it follows certain mechanism. Random variable can be a discrete it may be distinct values, countable. For example year is a discrete random variable. For example Mass it is a continuous random variable.

(Refer Slide Time: 04:02)

Probability Distributions

- The probability distribution function or probability density function (PDF) of a random variable X means the values taken by that random variable and their associated probabilities.
- PDF of a discrete r.v. (also known as PMF):
Example 1: Let the r.v. X be the number of heads obtained in two tosses of a coin.
Sample Space: {HH, HT, TH, TT}

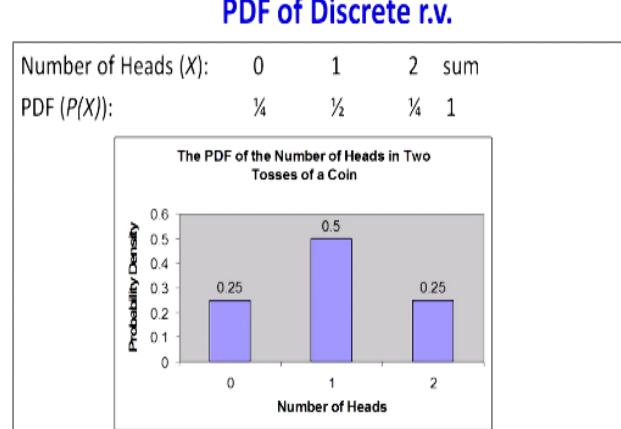


6

Then probability distributions, the probability distribution function or probability density function PDF of the random variable X means the values taken by the random variable and their associated probabilities if you make a relation between X and corresponding probabilities p of x or f of x, that if you plot that point that will form your distributions, so PDF of your discrete random variable also known as PMF probability mass function.

Example let the random variable X be the number of heads obtained in you 2 tosses of your coin. There are 2 possibilities when you toss 2 times 2 tosses, first toss you may get head, second toss you may get head then head tail, tail head, tail tail, so these are the sample space.

(Refer Slide Time: 04:53)



7

Probability density functions of your discrete random variable. Suppose we are tossing coin 2 times the probability of the 0 head is 1 by 4, the probability of getting one head is 1 by 2, the probability of getting 2 heads 1 by 4 some should be 1. See in the in the X axis, the random variable is taken 0 in Y axis corresponding probabilities marked. So, in X axis random variable and Y axis corresponding probability this is called the distributions.

(Refer Slide Time: 05:23)

Probability Distribution for the Random Variable X

A probability distribution for a discrete random variable X:

x	-8	-3	-1	0	1	4	6
$P(X = x)$	0.13	0.15	0.17	0.20	0.15	0.11	0.09

Find

a. $P(X \leq 0)$ 0.65

b. $P(-3 \leq X \leq 1)$ 0.67

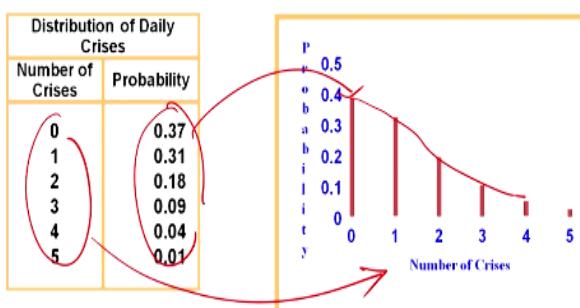


Now, probability distribution for a random variable X will do a small numerical problem, a probability distribution for discrete random variable X is given. So, X is given corresponding probability distribution is given. So, this is an empirical distribution. Suppose, if you want to know, what is the probability of $X \leq 0$. So, what you have to do? Wherever random variable X is 0 and less than equal to 0 you would add it.

For example $0.20 + 0.17 + 0.15 + 0.13$ we will get 0.65. Suppose, if you want to know the probability for the random variable - 3 to 1 - $\leq X \leq 1$. So, it would add - 3 to 1. $0.15 + 0.17 + 0.20 + 0.15$ and you add it will get 0.67.

(Refer Slide Time: 06:12)

Discrete Distribution -- Example



How to plot a discrete distribution? So number of crisis for example is taken the probability of happening that crises is also given for example, the probability of getting 0 crises is 0.37

for one crises 0.31 and so on. So, in X axis you mark the random variable in Y axis you plot the probability. When 0.37 it is a 1 this 1 this is that these are discrete these points, these points cannot be connected in x axis, this random variable has to come into the x axis, this probability has to go to Y axis.

For example, 0.37, this one here what will happen you cannot connect this line because it is a discrete, because you may have an $x = 1$ when $x = 2$, when $x = 1.5$ there is no value, if it is a discrete distribution, you cannot connect these points that is why it is called the discrete distributions.

(Refer Slide Time: 07:17)

Requirements for a Discrete Probability Function

- Probabilities are between 0 and 1, inclusively
- Total of all probabilities equals 1

$$0 \leq P(X) \leq 1 \quad \text{for all } X$$

$$\sum_{\text{over all } x} P(X) = 1$$



10

The requirement for the discrete probability density function, so probabilities are between 0 and 1 inclusively. Total of all probabilities equal to 1 and some are probability we have seen that already just 1

(Refer Slide Time: 07:30)

Cumulative Distribution Function

- The CDF of a random variable X (defined as $F(X)$) is a graph associating all possible values, or the range of possible values with $P(X \leq x)$.
- CDFs always lie between 0 and 1 i.e., $0 \leq F(X) \leq 1$, Where $F(X)$ is the CDF.



11

Next term will see cumulative distribution function. The cumulative distribution function of a random variable X defined as F of X is the graph associating all possible values are in the range of possible values with the P of $X \leq x$. Cumulative probability distribution function is just adding the probabilities. The CDF always lies between 0 to 1 that is $0 \leq F$ of x should be ≤ 1 F CDF cumulative density function.

(Refer Slide Time: 08:04)

The Expected Value of X

Let X be a discrete rv with set of possible values D and pmf $p(x)$. The *expected value* or *mean value* of X , denoted

$E(X)$ or μ_X , is

$$E(X) = \mu_X = \sum_{x \in D} x \cdot p(x)$$

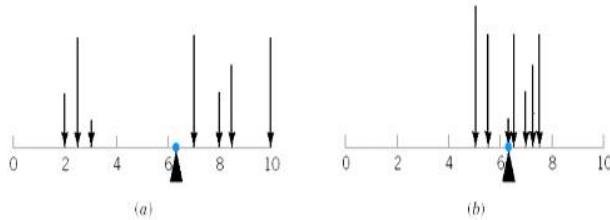


12

Then there is a very important property is the expected value of X . X be a discrete random variable with the set of possible values of D and pmf is $P(x)$. The expected value or mean value of X is denoted as, generally expect of x or $\mu_x = \sum x \cdot p(x)$. So, $\sum x \cdot p(x)$ is your expected value of X .

(Refer Slide Time: 08:32)

Mean and Variance of a Discrete Random Variable



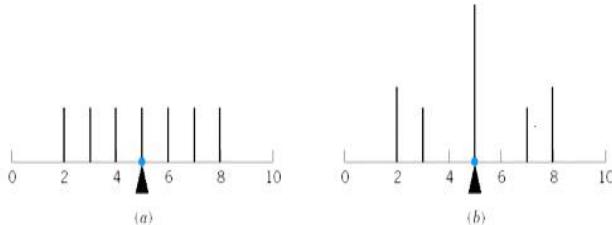
A probability distribution can be viewed as a loading with the mean equal to the balance point. Parts (a) and (b) illustrate equal means, but Part (a) illustrates a larger variance.



What is the meaning of this mean and variance of your discrete random variable is look at this there are picture (a) and picture (b) the left side, the mean is same for both the distribution, but look at the variance. The left side figure it shows that lot of variances that items figure it is less variances, the probability distribution can be viewed as, are viewed as you are loading with the mean equal to the balance point. So mean is nothing but it is like kind of a balance point for which the distribution lies. Part (a) and part (b) illustrate equal means, but part (a) illustrates a larger variance.

(Refer Slide Time: 09:12)

Mean and Variance of a Discrete Random Variable



The probability distribution illustrated in Parts (a) and (b) differ even though they have equal means and equal variances.



See the second case mean and variance of discrete random variable, the probability distribution illustrated in parts a and part b differs even though they have equal means and equal variances the shape of the distribution is different.

(Refer Slide Time: 09:25)

Example – Expected Value

- Use the data below to find out the expected number of credit cards that a customer to a retail outlet will possess.

$$x = \# \text{ credit cards}$$

x	P(x = X)
0	0.08
1	0.28
2	0.38
3	0.16
4	0.06
5	0.03
6	0.01

$$\begin{aligned}E(X) &= x_1 p_1 + x_2 p_2 + \dots + x_n p_n \\&= 0(.08) + 1(.28) + 2(.38) + 3(.16) \\&\quad + 4(.06) + 5(.03) + 6(.01) \\&= 1.97\end{aligned}$$

About 2 credit cards



15

Now, we will see how to find out an expected value use the data below to find out the expected number of credit cards that a customer to retail outlet will process. So X is a random variable. There is a how many number of credit cards customers having the P (x) equal to X corresponding probability. So Zero P (x) is .08. That means probability of a person having 0 credit card is 8 % probabilities a person to have for example, 6 credit cards is 1 %.

So how to find out the expected value multiplied by x and corresponding probability to submit. So, 0 (0.08) + 1 (0.28) + 2 (2.38), and so on, + 6 (.01) = 1.97. You can make them 2. That means the customers, they can have an average of 2 credit cards that was any customer if you take randomly average that customer can have 2 credit cards. Here an example of meaning of this what is mean.

(Refer Slide Time: 10:31)

The Variance and Standard Deviation

Let X have pmf $p(x)$, and expected value μ

Then the variance of X , denoted $V(X)$
(or σ_X^2 or σ^2), is

$$V(X) = \sum_D (x - \mu)^2 \cdot p(x) = E[(X - \mu)^2]$$

The standard deviation (SD) of X is $\sigma_X = \sqrt{\sigma_X^2}$



16

Now we will see how to find out the variance and standard deviation of an empirical distribution. Previously we have seen $\mu_x = \Sigma x \cdot p(x)$, I will see how to find out the variance of an empirical distribution. Let X have the pmf of p of x and the expected values μ . We know already the mean of an empirical distribution. Now we have to find out the variants of the empirical distribution, then the variance of X denoted as $V(X)$ or σ_x^2 or σ^2 .

The variance of $X = \Sigma (x - \mu)^2 \cdot p(x)$, variances can be denotes $E(X - \mu)^2$. The standard deviation is the square root of this.

(Refer Slide Time: 11:14)

The quiz scores for a particular student are given below:

22, 25, 20, 18, 12, 20, 24, 20, 20, 25, 24, 25, 18

Find the variance and standard deviation.

Value	12	18	20	22	24	25
Frequency	1	2	4	1	2	3
Probability	.08	.15	.31	.08	.15	.23

$$\mu = 21$$

$$V(X) = p_1(x_1 - \mu)^2 + p_2(x_2 - \mu)^2 + \dots + p_n(x_n - \mu)^2$$

$$\sigma = \sqrt{V(X)}$$



17

We will see you an example, a quiz scores for a particular student are given below 20, 25 and so on find the variance and standard deviation. So, before knowing the standard deviation, first you have to find out the mean because the mean is required. So, the mean if you add and

divided by corresponding elements, number of elements you will get 21. For example first we will construct a frequency distribution, you see 12 is repeated by 1 time, 18 is repeated by 2 times, 20 is repeated by 4 times, 25 is repeated by 3 times.

Then we have to find the probability, the probability is nothing but the relative frequency as I told you one definition of probability is relative frequency. So what is the cumulative frequency here first you have to find a total frequency $1 + 2 = 3$, $3 + 4 = 7$, $8, 10, 13$ there is a cumulative frequency. So, the probability here we are obtaining by using the concept of relative frequency. So the relative frequencies we are adding all the frequency that is a total. So 1 divided by corresponding sum of all frequencies 2 divided by some of frequencies. now the mu we can find out mu in another way also, we know that already we are done with this relative frequencies $\Sigma F \times M / \Sigma F$, how to find out the mean, sigma of expected value $X \cdot p(x)$, $12 \times 0.08 + 18 \times 0.15 + 20 \times 0.31 + 22 \times 0.08$,

$\Sigma x \cdot p(x) = 21$. One way you can add all the values you can divide by number of elements.

Otherwise from this empirical distribution, what x is given x is 12, 18, 20, probabilities given. So, if you want to know the mean x into p of x, now we are going to find out the variance.

(Refer Slide Tim: 13:40)

$$V(X) = .08(12-21)^2 + .15(18-21)^2 + .31(20-21)^2 \\ + .08(22-21)^2 + .15(24-21)^2 + .23(25-21)^2$$

$$V(X) = 13.25$$

$$\sigma = \sqrt{V(X)} = \sqrt{13.25} \approx 3.64$$

So p_1 variance = $.08 \cdot X_1$, Where $X_1 = (12 - \mu)^2$, p_2 , that is, $.15 \cdot X_2$ That is, $(18 - \mu)^2$ and so on when you add it you will get the variance and the mu take square root of will get this. You see that and going back $.08(12-21)^2 + .15(18-21)^2 + .31(20-21)^2$, when you simplify the variances 13.25 standard deviation is 3.64. So, what do I have done seen this problem, the data is given first you were constructed here empirical distributions, then we

will use the formula of needed variants to find out the mean and variance.

(Refer Slide Time: 14:13)

Shortcut Formula for Variance

$$V(X) = \sigma^2 = \left[\sum_D x^2 \cdot p(x) \right] - \mu^2 \quad \text{=} \quad E(x - \mu)^2 \\ = E(X^2) - [E(X)]^2$$



19

Another shortcut formula to find the variances is nothing but we the $E(x - \mu)^2$ for example already we seen $E(x - \mu)^2$, when you square it and simplify it, you will get this formula. So

$$V(X) = [\sum x^2 p(x)] - \mu^2 \\ = E(X^2) - [E(X)]^2$$

just you expand it will get this answer.

(Refer Slide Time: 14:48)

Mean of a Discrete Distribution

$$\mu = E(X) = \sum X \cdot P(X)$$

X	P(X)	X.P(X)
-1	.1	-.1
0	.2	.0
1	.4	.4
2	.2	.4
3	.1	.3
		1.0



20

So let us find out the meaning of your discrete distribution, the formula for finding the mean mu equal to expected value of X that is $X \cdot p(x)$. So X is given p(x) is given, multiply X and

$p(X)$ after doing that, when you sum the sum is 1. So the mean of this empirical distribution is 1.

(Refer Slide Time: 15:07)

Variance and Standard Deviation of a Discrete Distribution

$$\sigma^2 = \sum (X - \mu)^2 \cdot P(X) = 1.2$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.2} \approx 1.10$$

X	P(X)	$X - \mu$	$(X - \mu)^2$	$(X - \mu)^2 \cdot P(X)$
-1	1	-2	4	.4
0	.2	-1	1	.2
1	.4	0	0	.0
2	.2	1	1	.2
3	.1	2	4	.4
				1.2



21

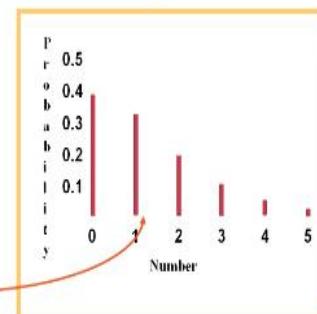
That is find out the variance and standard deviation of this empirical distribution that is a discrete distribution. So σ^2 , we know $(X - u)^2 \cdot p(x)$. So X is given p(x) is given first to find out $(X - u)$ then $(X - u)^2$, then multiply $(X - u)^2$ by $P(x)$ and sum it we are getting 1.2. So the variance is 1.2 you take square root the standard deviation is 1.10.

(Refer Slide Time: 15:37)

Mean of the Data Example

$$\mu = E(X) = \sum X \cdot P(X) = 1.15$$

X	P(X)	$X \cdot P(X)$
0	.37	.00
1	.31	.31
2	.18	.36
3	.09	.27
4	.04	.16
5	.01	.05
		1.15



22

Suppose then another distribution say X is given p of x is given X into p of X. So when you plot it the mean you see that the mean or mean not be exactly 1 or 2 or 3 mean value may in between 1 and 2. So mean value need not be discrete only the random variable is discrete here.

(Refer Slide Time: 16:00)

Properties of Expected Value

1. $E(b) = b$, b is a constant.
2. $E(X + Y) = E(X) + E(Y)$.
3. $E\left(\frac{X}{Y}\right) \neq \frac{E(X)}{E(Y)}$.
4. $E(XY) \neq E(X)E(Y)$ unless they are independent.
5. $E(aX) = aE(X)$, a constant.
6. $E(aX + b) = aE(X) + b$, a and b are constants.



23

Some of the very important properties of expected values suppose the expected value of a constant is constant only. When you want to multiply 2 random variable $E(X + Y)$. we can write $E(X) + E(Y)$. $E(X \setminus Y)$ is not division it is a conditional, it is kind of a conditional probability. So $E(X \setminus Y)$ need not be will not be equal to $E(X)$ divided by $E(Y)$ and same thing $E(XY)$ is not equal to $E(X)$ multiply $E(Y)$ unless they are independent.

If they are independent, you can write $E(XY)$ equals to $E(X)$ and $E(Y)$ otherwise we cannot written. So, if a random variable come along with the constant that constant can be removed out of this expected value. For example $E(aX)$ the a can be brought left side. So, $aE(X)$, here a is the constant. So, easy that it is in format $E(ax + b)$ that can be brought a left side. So, $aE(x)$, then when your expect b value constant this constant itself. It will become $aE(x + b)$ where a and b or constant.

(Refer Slide Time: 17:11)

Properties of Variance

1. $\text{Var}(\text{constant}) = 0$
2. If X and Y are two independent random variables, then
$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \text{ and}$$
$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$$
3. If b is a constant then $\text{Var}(b+X) = \text{Var}(X)$
4. If a is a constant then $\text{Var}(aX) = a^2\text{Var}(X)$
5. If a and b are constants then $\text{Var}(aX+b) = a^2\text{Var}(X)$
6. If X and Y are two independent random variables and a and b are constants then $\text{Var}(aX+bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$



Then properties of variances; variances of a constant is 0. If X and Y are 2 independent random variable, then $\text{Var}(X + Y) = \text{var}(X) + \text{var}(Y)$. $\text{Var}(X - Y) = \text{var}(X) + \text{var}(Y)$, it should be very carefully here, support there are 2 groups there are group 1 and group 2. If you want to know the difference in the variance, you too add their variances b a constant, then variances of $b + X$ because variances of b will become 0 it will become only variances of X .

If a is constant than variances of aX is because variances Ax the square term and you bring left side of the bracket would write a square and variances of X . There are proof is therefore this if a and b are constant than variances of $aX + b$ equal to a square variances of X , the variance of B will become 0. The answer is a square variances of X . If X and Y are 2 random variable and a and b are constant, then $\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y)$.

(Refer Slide Time: 18:24)

Covariance

Covariance: For two discrete random variables X and Y with $E(X) = \mu_x$ and $E(Y) = \mu_y$, the covariance between X and Y is defined as
$$\text{Cov}(XY) = \sigma_{xy} = E(X - \mu_x)(Y - \mu_y) = E(XY) - \mu_x \mu_y$$



Then covariance for 2 discrete random variable X and Y, $E(X) = \mu_x$ and $E(Y) = \mu_y$, then covariance between X and Y is the defined as covariance of XY equal to can be written as $\sigma_{xy} = E(X - \mu_x)(Y - \mu_y)$ and simplify it will get $E(XY) - \mu_x \cdot \mu_y$, that is a covariance.

(Refer Slide Time: 18:54)

Covariance

- In general, the covariance between two random variables can be positive or negative.
- If two random variables move in the same direction, then the covariance will be positive, if they move in the opposite direction the covariance will be negative.

Properties:

1. If X and Y are independent random variables, their covariance is zero. Since $E(XY) = E(X)E(Y)$
2. $\text{Cov}(XX) = \text{Var}(X)$
3. $\text{Cov}(YY) = \text{Var}(Y)$



26

In general, the covariance between 2 random variable can be positive or negative. If random variables move in the same direction, then the covariance will be positive. If they move in the opposite direction, the covariance will be negative. Properties of covariance, if X and Y are independent random variables their covariance is 0. Since $E(XY) = E(X)E(Y)$ is independent covariance there would not be any covariance. Covariance (XX) is variance of X. Similarly, covariance of YY is simply variance of Y.

(Refer Slide Time: 19:31)

Correlation Coefficient

- The covariance tells the sign but not the magnitude about how strongly the variables are positively or negatively related. The correlation coefficient provides such measure of how strongly the variables are related to each other.
- For two random variables X and Y with $E(X) = \mu_x$ and $E(Y) = \mu_y$, the correlation coefficient is defined as

$$\rho_{xy} = \frac{\text{Cov}(XY)}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$



27

Then correlation coefficient, the covariance tells the sign but not the magnitude about how strongly the variables are positively or negatively related. The correlation coefficient provides such measures of how strongly the variables are related to each other. The variance is only giving the direction not the magnitude, but the correlation it is giving the magnitude for 2 random variables X and Y, $E(X) = \mu_x$ and $E(Y) = \mu_y$. The correlation coefficient is defined as covariance of X Y divided by σ_x and σ_y .

(Refer Slide Time: 20:06)

Some Special Distributions

- Discrete
 - Binomial
 - Poisson
 - Hyper geometric
- Continuous
 - Uniform
 - Exponential
 - Normal



28

Dear students now are going to some special distributions will study some special distribution in a discrete category and continuous category. The discrete will study about the binomial distribution and Poisson distribution and Hyper geometric distribution. Continuous category which will study we are going to study uniform exponential and normal. In this class I will explain the theory and corresponding its parameters outer end of the class will use Python to find out various parameters various mean and variance of your distributions and corresponding probabilities in the practical class.

(Refer Slide Time: 20:46)

Binomial Distribution

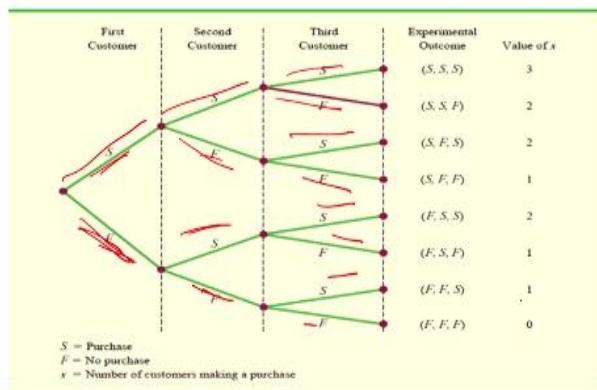
- Let us consider the purchase decisions of the next three customers who enter a store.
- On the basis of past experience, the store manager estimates the probability that any one customer will make a purchase is .30.
- What is the probability that two of the next three customers will make a purchase?

29

First one is the binomial distribution. Let us consider an example to explain the concept of binomial distribution. Let us consider the purchase decision of the next 3 customers who enter in store there are 3 customers going enter the store and the basis of past experience, the store manager estimates that the probability that any one customer will make your purchase is 0.30. What is the probability that 2 of the next 3 customers will make a purchase?

(Refer Slide Time: 21:18)

Tree diagram for the Martin clothing store problem



30

Now look at this the tree diagram the first customer there is a 2 possibility, S is the purchase F is no purchase, X is the number of customers making purchase. So we will see that is the end here. Now, what is happening the first customer he can purchase or not purchased second customer different possibilities, third customer different possibilities. Now we look at the experimental outcome, this possibility, look at this possibility, success success success.

Look at this possibility success success failure look at this personal success failure Success, then success failure failure, failure success success, failure success failure, failure failure success, failure failure failure. So, we have written all possibilities. Now, the question is out of 3 customers what is the probability that 2 customers will make a purchase? What is the meaning SSS all 3 customer have purchased.

So value of x equal to 3 random variable second case to customer how purchased third customer did not buy. So, here x is 2 because here the X is taken the number of customers making purchase the first possibility $x = 3$, the second possibility is 2, the third possibility 2, the fourth possibility is 1,.. 2, 1, 1, 0. Now the question is, what is the probability that 2 out of 3 customers will make a purchase, See that? There is a possibility.

(Refer Slide Time: 23:00)

Trial Outcomes

Trial Outcomes			Experimental Outcome	Probability of Experimental Outcome
1st Customer Purchase	2nd Customer Purchase	3rd Customer No purchase		
Purchase	Purchase	No purchase	(S, S, F)	$pp(1 - p) = p^2(1 - p)$ = (.30)^2(.70) = .063
Purchase	No purchase	Purchase	(S, F, S)	$p(1 - p)p = p^2(1 - p)$ = (.30)^2(.70) = .063
No purchase	Purchase	Purchase	(F, S, S)	$(1 - p)pp = p^2(1 - p)$ = (.30)^2(.70) = .063

$$3C_2 = 3$$



31

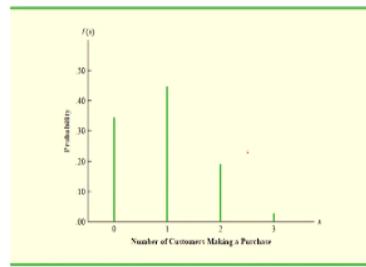
The first customer: it is the possibility, the SSF, SFS FSS what is the probability of success is p,p and $1 - p$ we know p is .3. So $0.3^2 \times 0.7 = .063$. For second category also we are getting p , first success p failure $1 - p$ again success is p . So, p^2 multiplied by $(1 - p)$. so $0.3^2 \times 0.7$ equal to $.063$ then third possibility failure, $1 - p$ success success p, p , So, $p^2 \times (1 - p) = .063$, $.063$, $.063$.

Now, here we actually do here the possibilities 3C_2 because the question is asked out of 3 customers, what is the probability that 2 customer will buy? So it is a 3C_2 , the value of 3C_2 is 3. That is why 1, 2, there are 3 possibilities when you go back when you go back how many 3 is there 1 2 3 possibilities there is the meaning of value 3C_2 .

(Refer Slide Time: 24:20)

Graphical representation of the probability distribution for the number of customers making a purchase

x	P(x)
0	$0.7 \times 0.7 \times 0.7 = 0.343$
1	$0.3 \times 0.7 \times 0.7 + 0.7 \times 0.3 \times 0.7 + 0.7 \times 0.7 \times 0.3 = 0.441$
2	0.189
3	0.027



32

Now we will find out the probability if $x = 0$ that mean nobody is buying.

(Refer Slide Time: 24:33)

$$\begin{aligned} S^2 &= \sum (x - \bar{x})^2 / n-1 \\ \text{cov}(x,y) &= \sum (x - \bar{x})(y - \bar{y}) / n-1 \\ \text{correlation coefficient} &= \frac{\text{cov}(x,y)}{\sigma_x \cdot \sigma_y} \\ m - \text{slope of a regression Eqn} &= \frac{\text{cov}(x,y)}{\text{Var } x} = m = \text{slope of Regr.} \end{aligned}$$



28

Students we have studied so far variance, covariance, correlation coefficient, just how to make it in relation. For example, we know the variance formula, variance equal to $\sum(x - x \bar{})^2$. So variances for 1 variable, suppose if you want to know 2 variable if you want to know for 2 variable, this variance will be called this covariance.

So covariance is $\sum(x - x \bar{})(Y - Y \bar{})$ variances divided by $n - 1$. So, $\sum(x - x \bar{})(Y - Y \bar{})$, here also $n - 1$. Variance, covariance, next one correlation coefficient is covariance x, y divided by standard deviation x, standard deviation y. Now, you see that this is a variances, this is a covariance, correlation coefficient. So, for correlation coefficient when you divide covariance to a corresponding standard deviation you will get correlation coefficient.

Next we will say ‘m’ that is called slope of the regression, slope of a regression equation. So, there is nothing but you were covariance(x, y) divided by variances of x, the first one is variance, covariance, correlation coefficient and slope of regression equation, you see that all are having some relationship for the variance is only for one variable, what is the meaning of variance?

How each value is away from its mean that deviation square of the deviation, then some of the deviation, then the mean value of that some of the deviation it will give you the variance for covariance there are 2 variables there how each variable is moving away from its own mean. So, $\Sigma(x - x \bar{ }) (y - y \bar{ })$ divided by $n - 1$. If you want to know correlation coefficient, that covariance is divided by its corresponding standard deviation look at correlation coefficient.

If you want to know slope and regression equation, if you divide covariance of x, y divided by corresponding variances of x you will get m that is nothing but slope of the regression equation. Dear students so far we have seen what is the need for studying the distribution?, then we have seen how to construct a discrete probability distribution after constructing how to find out the mean and variance of a discrete distribution then we have seen the properties of expected value.

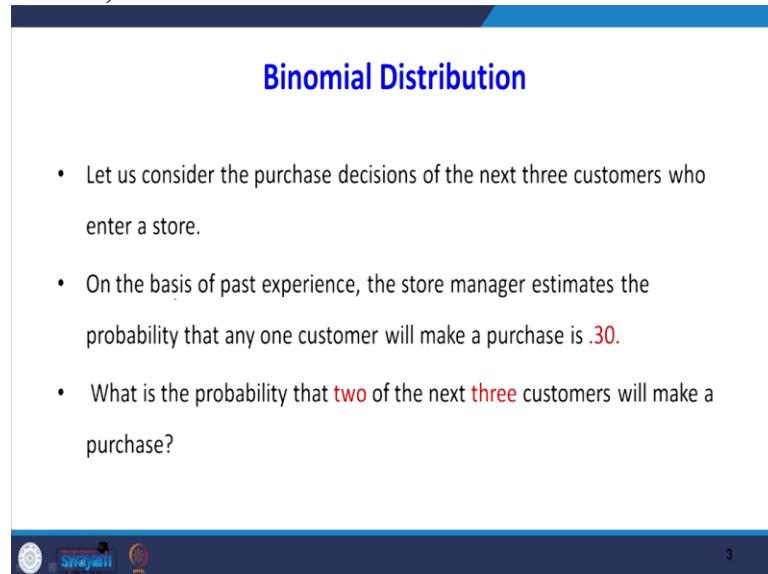
Next we have seen the properties of the variance. Then we have seen how this mean, variances, covariance are interrelated. The next class we will continue some discrete distributions and some continuous distribution in detail. Thank you

Data Analytics with Python
Prof. Ramesh Anbanadham
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture - 09
Probability Distribution – II

Dear students in this lecture, we are going to see some of the discrete distributions and continuous distributions, the discrete we are going to study the binomial poisson hypergeometric in continuous that where we going to study uniform exponential and normal, first you will see, what is the application of this Binomial Distribution?

(Refer Slide Time: 00:48)



Binomial Distribution

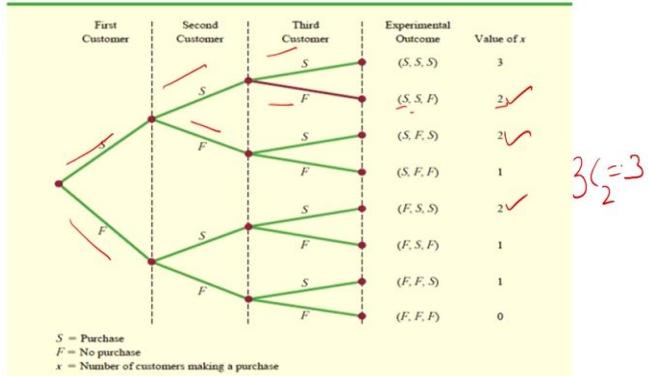
- Let us consider the purchase decisions of the next three customers who enter a store.
- On the basis of past experience, the store manager estimates the probability that any one customer will make a purchase is .30.
- What is the probability that two of the next three customers will make a purchase?

Otherwise what is the need of this binomial distribution. We will take one practical example that I will solve manually with the help of our concept of probability then I will tell you how the concept of binomial distribution will help us to solve this problem quick way, but just consider the purchase decision of the next 3 customers who enter your store and the basis of past experience.

The store manager estimates the probability that any one customer will make your purchase is 0.30. Now, what is the probability that 2 of the next 3 customers will make a purchase?

(Refer Slide Time: 01:29)

Tree diagram for the Martin clothing store problem



4

That problem is drawn in the form of a tree diagram the first customer, second customer, third customer when the first customer comes, there are 2 possibility he can go for purchase or not purchase, purchase mention as say S success failure. The second customer also may go for purchase, not purchase third customer purchase, not purchase. Look at the first possibility the First customer is purchasing, so, success, success, success that is S, S, S.

If you take X look at the bottom of this slide, if X equals the number of customers making a purchase, so, in this category, all 3 customers all 3 are going to buy look at the second possibility success, success failure. So, in this out of 3 customers 2 customers S S going to purchase like this we have displayed all possibilities. Now, the problem is X what is asked is 2 out of 3 customers what they have purchased.

So, number of customers making a purchases how many possibilities are there 2 customers 1, 2, 3 how we got this 3 is nothing but your 3C_2 the 3C_2 is out of 3, what is the probability? What is the chance that 2 customers will buy there will be a possibility of 3C_2 , so the 3C_2 is 3.

(Refer Slide Time: 02:57)

Trial Outcomes

Trial Outcomes				Probability of Experimental Outcome
1st Customer Purchase	2nd Customer Purchase	3rd Customer No purchase	Experimental Outcome	
Purchase	Purchase	No purchase	(S, S, F)	$pp(1-p) = p^2(1-p)$ $= (.30)^2(.70) = .063$
Purchase	No purchase	Purchase	(S, F, S)	$p(1-p)p = p^2(1-p)$ $= (.30)^2(.70) = .063$
No purchase	Purchase	Purchase	(F, S, S)	$(1-p)pp = p^2(1-p)$ $= (.30)^2(.70) = .063$

$P(X=2) = \binom{3}{2} p^2 q^{3-2}$



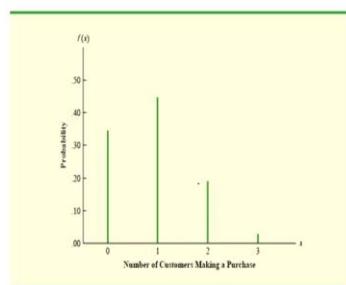
5

Now, we will see the first customer all possibilities purchase purchase no purchase success success, so, success we know, we are calling it is a p failure is $1 - p$. So, it is a $p.p$ into $1 - p$, p is $(0.30)^2$ multiply by 0.70, it is a 0.63. For second chance also first customer may buy second customer did not buy third customer also buy, so, it is success failure success. So, $p 1 - p$ into p again p^2 . $(1 - p) = 0.063$, third choice, no purchase purchase purchase FSS, so $1 - p$, p,p so it is a $p^2 . (1 - p) = 0.63$.

(Refer Slide Time: 03:48)

Graphical representation of the probability distribution for the number of customers making a purchase

x	P(x)
0	$0.7 \times 0.7 \times 0.7 = 0.343$
1	$0.3 \times 0.7 \times 0.7 + 0.7 \times 0.3 \times 0.7 + 0.7 \times 0.7 \times 0.3 = 0.441$
2	0.189
3	0.027



6

If I plot this one see x possibility, x possibility this possibility is see this possibility, when $x = 0$, that means, what is the charge that all 3 customers will not buy? That is this case 0.7, 0.7, 0.7. If I say $x = 1$, what is the probability that 1 customer will buy, if $x = 2$? What is the probability 2

customers will come, 2 customers will buy with $x = 3$? What is the probability that 3 customers will buy? So, we can do with the concept of probability manually like this.

So, the question is asked is this one? What is the probability that 2 customers will buy out of 3. So, this is your probability distributions. So, this we can do manually but it will take a lot of time. So, here with the help of your binomial distributions, you can get right this all possibilities going back this way nC_x So,

$P(X=2) = {}^nC_x \cdot p^x \cdot q^{n-x}$, where $n = 3$ where $x = 2$ p is the probability success q = probability of failure.

So, this nC_x will give you different combinations of that event to happen. So, this p^x , see that always here $p^2 q^{n-2}$. So, this will give you the probability. So, that is the purpose of this binomial distributions.

(Refer Slide Time: 05:35)

Binomial Distribution- Assumptions

- Experiment involves n identical trials
- Each trial has exactly two possible outcomes: success and failure
- Each trial is independent of the previous trials
- p is the probability of a success on any one trial
 $q = (1-p)$ is the probability of a failure on any one trial
- p and q are constant throughout the experiment
- X is the number of successes in the n trials

What is the property of this binomial distribution experiment involves in identical trials, each trial has exactly 2 possible outcomes success, failures. Each trial is independent of the previous trial for example, the previous case customer one may buy or may not buy. That is not depending upon the for example for the second customer is intension of purchase is not based on the previous customer.

So, p is the probability of success on any one trial, the q is probability of failure on any one trial. So, p and q are constant throughout the experiment, this is important assumption. So, x is the number of success in the end trials, there the previous case we are seeing x = 2.

(Refer Slide Time: 06:23)

Binomial Distribution

- Probability function $P(X) = \frac{n!}{X!(n-X)!} p^X q^{n-X}$ for $0 \leq X \leq n$
- Mean value $\mu = n \cdot p$
- Variance and standard deviation $\sigma^2 = n \cdot p \cdot q$
 $\sigma = \sqrt{\sigma^2} = \sqrt{n \cdot p \cdot q}$

8

So the probability function as I told you can written as

$${}^nC_x = (n!/(X!(n-X)!)) \cdot p^x \cdot q^{n-x}$$

the mean of this distribution is $\mu = n.p$ directly we are going to answered. So, you can find out I think but using the concept of expected value, when you multiply this P (X) with the x, $\sigma_x \cdot P(x)$, when you simplify will get this formula. Similarly, variance directly we are going to use only the result variance = nPq

standard deviation = root of nPq .

(Refer Slide Time: 07:03)

		Binomial Table									
		SELECTED VALUES FROM THE BINOMIAL PROBABILITY TABLE									
		EXAMPLE: $n = 10, x = 3, p = .40; f(3) = .2150$									
<i>n</i>	<i>x</i>	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
9	0	6302	3874	2316	1342	0751	0404	0207	0091	0046	0020
	1	2985	3874	3679	3020	2253	1556	1094	0695	0319	0176
	2	0629	1722	2597	3020	3003	2668	2162	1412	1110	0703
	3	0077	6446	1069	1762	2336	2668	2716	2508	2119	1641
	4	0006	0074	0233	0661	1168	1715	2194	2508	2800	2461
	5	0000	0008	0050	0165	0389	0735	1181	1672	2128	2461
	6	0000	0001	0006	0023	0087	0210	0454	0743	1160	1641
	7	0000	0000	0000	0005	0012	0039	0098	0212	0407	0703
	8	0000	0000	0000	0000	0001	0004	0013	0035	0083	0176
	9	0000	0000	0000	0000	0000	0000	0001	0003	0008	0020
10	0	5987	3487	1969	1074	0563	0282	0135	0060	0025	0010
	1	3151	3874	3474	2684	1877	1211	0725	0403	0207	0098
	2	0746	1937	2759	3020	2816	2335	1757	1209	0763	0439
	3	0105	0574	1298	2013	2503	2668	2522	2150	1665	1172
	4	0010	0112	0401	0881	1460	2001	2377	2508	2384	2051
	5	0001	0015	0085	0264	0584	0129	0536	0207	0340	0261
	6	0000	0001	0012	0055	0162	0368	0689	1115	1596	2051
	7	0000	0000	0001	0008	0031	0090	0212	0425	0746	1172
	8	0000	0000	0000	0001	0004	0014	0043	0106	0229	0439
	9	0000	0000	0000	0000	0000	0001	0005	0016	0042	0089
	10	0000	0000	0000	0000	0000	0000	0001	0003	0010	



9

Sometime there is ready made table is there to find out binomial probability values for example, $n = 10$ and n is number of trials, $x = 3$, $p = 0.04$, $n = 10$, $x = 3$ that value is this one 0.150. So, the tables are provided you do not use your calculator it may take more time.

(Refer Slide Time: 07:34)

Mean and Variance

- Suppose that for the next month the Clothing Store forecasts 1000 customers will enter the store.
- What is the expected number of customers who will make a purchase?
- The answer is $\mu = np = (1000)(.3) = 300$.
- For the next 1000 customers entering the store, the variance and standard deviation for the number of customers who will make a purchase are

$$\sigma^2 = np(1-p) = 1000(.3)(.7) = 210$$

$$\sigma = \sqrt{210} = 14.49$$



10

Now we will find out mean and variance suppose, for the next month the clothing store forecast 1000 customers will enter the store the previous problem their 1000 customers. So, what is the expected number of customers who will make the purchase out of 1000. So, the answer is $\mu = n p$ so in this 1000 probability of success point 3, so it is equal to 300. So, that is out of 1000 customers, there is a chance that 300 customers will make the purchase.

For the next 1000 customers entering the store the variance and standard deviation of the number of customers who will make the purchase can you written as npq . So, n is given p is given $1 - p = 0.210$ the standard deviation is 14.49.

(Refer Slide Time: 08:24)

Poisson Distribution

- Describes discrete occurrences over a continuum or interval
- A discrete distribution
- Describes rare events
- Each occurrence is independent any other occurrences.
- The number of occurrences in each interval can vary from zero to infinity.
- The expected number of occurrences must hold constant throughout the experiment.



11

Then will go to next distribution Poisson distribution describes discrete occurrences over continuous or interval. Generally Poisson distribution is for rare events, it is an discrete distribution. X can have only few discrete values. Yes, it describes the rare events, each occurrence is independent of any other occurrences. The number of occurrences in each interval can vary from 0 to infinity. The expected number of occurrences must hold constant throughout the experiment. These are the assumption of Poisson distribution.

(Refer Slide Time: 09:09)

Poisson Distribution: Applications

- **Arrivals at queuing systems**
 - airports -- people, airplanes, automobiles, baggage
 - banks -- people, automobiles, loan applications
 - computer file servers -- read and write operations
- **Defects in manufactured goods**
 - number of defects per 1,000 feet of extruded copper wire
 - number of blemishes per square foot of painted surface
 - number of errors per typed page



12

Some of the examples are application of Poisson distribution. So, arrival at the queuing system follow Poisson distribution, any airport people may arrive, airplane may arrive automobile may arrive, baggage may arrive that follow Poisson distribution. The banks people automobiles loan applications this arrival pattern follow Poisson distribution in computer file services, read and write operations will follow Poisson distribution

The defects in manufacturing goods can be considered an example of Poisson distribution. For example, number of defects per 1000 feet of extruded copper wire see that the n is very large. The probability of success is very low. Another example of Poisson distribution is number of blemishes per square foot of painted surface blemishes kind of a defect in paint number of errors per typed page, these are the example of your Poisson distributions.

(Refer Slide Time: 10:15)

Poisson Distribution

- Probability function

$$P(X) = \frac{\lambda^X e^{-\lambda}}{X!} \text{ for } X = 0, 1, 2, 3, \dots$$

where:

λ = long-run average

$e = 2.718282\dots$ (the base of natural logarithms)

Mean value

$$\lambda$$

Variance

$$\lambda$$

Standard deviation

$$\sqrt{\lambda}$$



13

So, the probability function for Poisson distribution = $(\lambda^X e^{-\lambda})/X!$

Where X can take only discrete value, the lambda is mean that is a long run average the value of e is 2.71. The mean of this Poisson distribution is lambda variance of the Poisson distribution lambda standard deviation of Poisson distribution is root of lambda. So, this distribution called the univariate distribution.

Because it has unique parameter that means, it is a special property of your Poisson distribution where mean and variance is same. So, that is mean and variances. Same that is lambda.

(Refer Slide Time: 10:55)

Poisson Distribution: Example

$\lambda = 3.2 \text{ customers/4 minutes}$

$X = 10 \text{ customers/8 minutes}$

Adjusted λ

$\lambda = 6.4 \text{ customers/8 minutes}$

$$P(X) = \frac{\lambda^X e^{-\lambda}}{X!}$$

$$P(X=10) = \frac{6.4^{10} e^{-6.4}}{10!} = 0.0528$$

$\lambda = 3.2 \text{ customers/4 minutes}$

$X = 6 \text{ customers/8 minutes}$

Adjusted λ

$\lambda = 6.4 \text{ customers/8 minutes}$

$$P(X) = \frac{\lambda^X e^{-\lambda}}{X!}$$

$$P(X=6) = \frac{6.4^6 e^{-6.4}}{6!} = 0.1586$$



14

And another caution while using Poisson distribution is the unit of this lambda and X should be same, for example, lambda equal to 3.2 customers for 4 minutes probability of around 10 customers in 8 minutes. So, you have to adjust the lambda, how will you adjust, you multiply by 2, so it will be 6.4 into 8. Now, the unit of X and lambda is same, then you can use your probability function, $P(X) = (\lambda^X e^{-\lambda})/X!$ and substitute $X = 10$.

We are getting 0.025. See the second one, lambda = 3.2 customers by 4 minutes, X = 6 customers for 8 minutes. So, you have to multiply lambda. So, it will be 6.4 in 8 then P of X will get the answer. So, what is the point here is the unit of lambda and X should be same.

(Refer Slide Time: 11:49)

Poisson Probability Table											
Example: $\mu = 10, x = 5; f(5) = .0378$											
x	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10	μ
0	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000
1	0.010	0.009	0.009	0.008	0.007	0.006	0.005	0.005	0.005	0.005	0.005
2	0.046	0.043	0.040	0.037	0.034	0.031	0.029	0.027	0.025	0.023	0.023
3	0.140	0.131	0.123	0.115	0.107	0.100	0.093	0.087	0.081	0.076	0.076
4	0.319	0.302	0.285	0.269	0.254	0.240	0.226	0.213	0.201	0.189	0.189
5	0.581	0.555	0.530	0.506	0.483	0.460	0.439	0.418	0.398	0.378	0.378
6	0.881	0.851	0.822	0.793	0.764	0.736	0.709	0.682	0.656	0.631	0.631
7	1.145	1.118	1.091	1.064	1.037	1.010	0.982	0.955	0.928	0.901	0.891
8	1.302	1.286	1.269	1.251	1.232	1.212	1.191	1.170	1.148	1.126	1.126
9	1.317	1.315	1.311	1.306	1.300	1.293	1.284	1.274	1.263	1.251	1.251
10	1.198	1.210	1.219	1.228	1.235	1.241	1.245	1.249	1.250	1.251	1.251
11	0.991	1.012	1.031	1.049	1.067	1.083	1.098	1.112	1.125	1.137	1.137
12	0.752	0.776	0.799	0.822	0.844	0.866	0.888	0.908	0.928	0.948	0.948
13	0.526	0.545	0.572	0.594	0.617	0.640	0.662	0.685	0.707	0.729	0.729
14	0.342	0.361	0.380	0.399	0.419	0.439	0.459	0.479	0.500	0.521	0.521
15	0.208	0.221	0.235	0.250	0.265	0.281	0.297	0.313	0.330	0.347	0.347
16	0.118	0.127	0.137	0.147	0.157	0.168	0.180	0.192	0.204	0.217	0.217
17	0.063	0.069	0.075	0.081	0.088	0.095	0.103	0.111	0.119	0.128	0.128
18	0.032	0.035	0.039	0.042	0.046	0.051	0.055	0.060	0.065	0.071	0.071
19	0.015	0.017	0.019	0.021	0.023	0.026	0.028	0.031	0.034	0.037	0.037
20	0.007	0.008	0.009	0.010	0.011	0.012	0.014	0.015	0.017	0.019	0.019
21	0.003	0.003	0.004	0.004	0.005	0.006	0.006	0.007	0.008	0.009	0.009
22	0.001	0.001	0.002	0.002	0.002	0.002	0.003	0.003	0.004	0.004	0.004
23	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.002	0.002
24	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.001	0.001

15

Here also there is a probability Poisson probability table is available. So, when $\mu = 10 X = 5$, you see in the column shows mu. So, $\mu = 10$ here, $X = 5$ is this value. So, this value is 0.378. So, you cannot, you need not to use your calculator directly you can read it from the table. But we are going to use these things in your practical class we are going to find out probabilities with the help of Python.

(Refer Slide Time: 12:23)

The Hypergeometric Distribution

- The **binomial distribution** is applicable when selecting from a finite population with replacement or from an infinite population without replacement.
- The hypergeometric distribution is applicable when selecting from a **finite population without replacement**.



We will go for hypergeometric distribution, see the binomial distribution is applicable, when selecting from finite population with replacement or for an infinite population without replacement. So, whenever the concept of without replacement comes, then we have to think of using hypergeometric distribution. The hypergeometric distribution is applicable when selecting from your finite population without replacement.

(Refer Slide Time: 12:52)

Hyper Geometric Distribution

- Sampling without replacement from a finite population
- The number of objects in the population is denoted N.
- Each trial has exactly two possible outcomes, success and failure.
- Trials are not independent
- X is the number of successes in the n trials
- The binomial is an acceptable approximation, if $N/10 > n$ Otherwise it is not.



17

The properties of hypergeometric distribution, so, sampling without replacement from your finite population then you should go for hypergeometric distribution. The number of objects in the population is denoted as N, each trial has exactly 2 possible outcomes, success or failure. Similar to binomial distribution trials are not independent, this is one different property, when compared

to binomial. In the binomial distributions, since we are going with replacement, the trials are independent; here we are going without replacement.

So, the trials are not independent, it is dependent. Then, that means, the P will not be fixed, the probability of P will not be fixed, every time we will get different answer for the P. X is the number of success in the n trials. The binomial is acceptable approximation $N/10 \geq n$, otherwise, it is not.

(Refer Slide Time: 13:53)

Hypergeometric Distribution

- Probability function
 - N is population size ✓
 - n is sample size ✓
 - A is number of successes in population ✓
 - x is number of successes in sample ✓

$$P(x) = \frac{\binom{A}{x} \binom{N-A}{n-x}}{\binom{N}{n}}$$

$$\mu = \frac{A \cdot n}{N}$$

- Mean Value
- Variance and standard deviation

$$\sigma^2 = \frac{A(N-A)n(N-n)}{N^2(N-1)}$$

$$\sigma = \sqrt{\sigma^2}$$

18

That we will see the probability function with discrete values use probability mass function, if it is a continuous will use PDF probability density function. So, the probability mass function is $P(x) = \binom{A}{x} \binom{N-A}{n-x} / \binom{N}{n}$. So, here see the capital letters represent for the population. So N is for population size, A is the number of success in the population, small letter, n for the sample size, the x is number of success in the sample.

So, $P(x) = \binom{A}{x} \binom{N-A}{n-x} / \binom{N}{n}$. The mean value = $A \cdot n / N$. The variance and of hypergeometric distribution is $A(N-A)n(N-n) / N^2(N-1)$, root of variance is the formula for standard deviation.

(Refer Slide Time: 14:57)

The Hypergeometric Distribution Example

- Different computers are checked from 10 in the department. 4 of the 10 computers have illegal software loaded.
- What is the probability that 2 of the 3 selected computers have illegal software loaded?
- So, N = 10, n = 3, A = 4, X = 2

$$P(X = 2) = \frac{\binom{A}{X} \binom{N - A}{n - X}}{\binom{N}{n}} = \frac{\binom{4}{2} \binom{6}{1}}{\binom{10}{3}} = \frac{(6)(6)}{120} = 0.3$$

$\textcircled{1}$ $\textcircled{2}$
 $N = 10$ $n = 3$
 $A = 4$ $x = ?$

- The probability that 2 of the 3 selected computers have illegal software loaded is .30, or 30%.



Will see one problem here, different computers are checked from 10 in the department, 4 out of 10 computers have illegal software loaded. What is the probability that 2 of the 3 selected computers will have illegal software loaded. By looking at the problem we are feeling that it is a finite population. Whenever there is a finite population, we should think of hypergeometric distribution.

So, N is given that N is your population size, A is 4, because we know 4 is out of 10, 4 computers are illegal software. Because we know the how much illegal software is installed in the population A = 4. Then, what is the probability that x is 2 of the 3 selected the computers will have illegal software loaded. So, x = 2, here n is 3 there are 2 things is there, one is for the population and population N = 10, A = 4.

For the sample n = 3, then what is the probability of that 2 out of 3 selected computers will have illegal software loaded. So now substituted, everything is given, so we will get 0.3. So what is the meaning of this 0.3, the probability that 2 of the 3 selected computers will have illegal software loaded is 30%.

(Refer Slide Time: 16:38)

Continuous Probability Distributions

- A continuous random variable is a variable that can assume any value on a continuum (can assume an uncountable number of values)
 - thickness of an item
 - time required to complete a task
 - temperature of a solution
 - height
- These can potentially take on any value, depending only on the ability to measure precisely and accurately.



Then we will go to the continuous probability distributions. A continuous random variable is a variable that can assume any value in the continuum. That is, can assume an uncountable number of values. Thickness of an item, time required to complete a task, temperature of a solution, height, these are example of continuous random variable. This can potentially take any value depending only on the ability to measure precisely and accurately.

The Uniform Distribution

- The uniform distribution is a probability distribution that has equal probabilities for all possible outcomes of the random variable
- Because of its shape it is also called a rectangular distribution

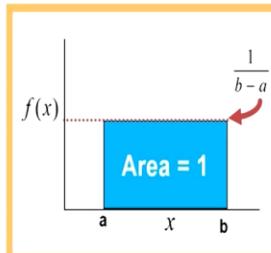


First we will see uniform distribution. The uniform distribution is the probability distribution that has equal probabilities for all possible outcomes of a random variable. That is the probability of random numbers also. When we say random numbers, the probability of choosing any number is same, because of its shape, it is also called a rectangular distribution.

(Refer Slide Time: 17:31)

Uniform Distribution

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{for all other values} \end{cases}$$



Look at this, Uniform Distribution. The uniform distribution is defendant interval, $a \leq x \leq b$. So the probability function is $1/ (b - a)$, where 'b' is upper limit, 'a' is the lower limit. In other interval, the value of $f(x)$ is 0, the area = 1.

(Refer Slide Time: 17:51)

Uniform Distribution: Mean and Standard Deviation

Mean

$$\mu = \frac{a + b}{2}$$

Standard Deviation

$$\sigma = \frac{b-a}{\sqrt{12}}$$



The mean of a uniform distribution is $\mu = (a + b)/2$,

Standard deviation = $(b - a) / \sqrt{12}$. These very standard result, the derivation of this one, you can refer some textbook.

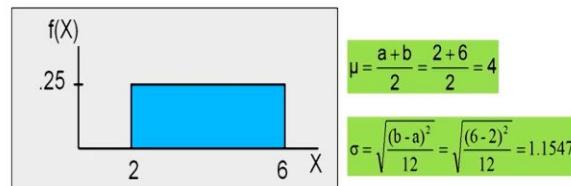
(Refer Slide Time: 18:07)

The Uniform Distribution

345

Example: Uniform probability distribution over the range $2 \leq X \leq 6$:

$$f(X) = \frac{1}{6 - 2} = .25 \text{ for } 2 \leq X \leq 6$$



Now, we will see some problem using uniform distribution, suppose uniform probability distribution over the range is defined this way. $2 \leq x \leq 6$. So if you want to know $f(x) = 1 / (b - a)$, where b is 6, a is 2, so, $(1 / (6 - 2)) = 0.25$. What is the meaning of this 0.25 means, in between 2 and 6, you can select any random variable, maybe 3, you may select 3 or 4 or 5, the probability of choosing 3 or 4 5 is, 0.25. So, the mean = $(a + b) / 2$,

where a is 2 b is 6

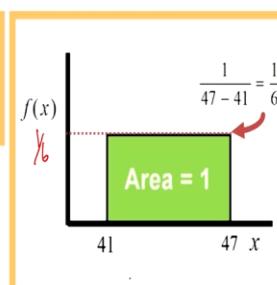
$= (8 / 2) = 4$. Similarly,

standard deviation = $(b - a) / \sqrt{12}$, that is 1.17.

(Refer Slide Time: 18:56)

Uniform Distribution Example

$$f(x) = \begin{cases} \frac{1}{47-41} & \text{for } 41 \leq x \leq 47 \\ 0 & \text{for all other values} \end{cases}$$



We will see another example, suppose a random variable is defined between 41 to 47. First, we have to find out probability density function 1 divided by b - a, so $47 - 41$, so that is $1 / 6$. So $1 / 6$ the height of this rectangle of distribution is $1 / x$ this value is your $1 / 6$. The lower limit 41 upper limit 47.

(Refer Slide Time: 19:21)

Uniform Distribution: Mean and Standard Deviation

Mean	Mean
$\mu = \frac{a + b}{2}$	$\mu = \frac{41 + 47}{2} = \frac{88}{2} = 44$
Standard Deviation	Standard Deviation
$\sigma = \frac{b - a}{\sqrt{12}}$	$\sigma = \frac{47 - 41}{\sqrt{12}} = \frac{6}{3.464} = 1.732$

So, the mean of the distribution is $(41 + 47)$ divided by 2

$$= 88 \text{ divided by } 2 = 44$$

$$\text{standard deviation } (b - a) / \sqrt{12}. \text{ So, } (47 - 41) / \sqrt{12}$$

$$= (47 - 41) / 3.464$$

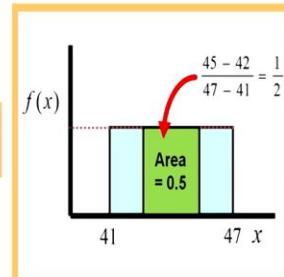
$$= 1.732.$$

(Refer Slide Time: 19:37)

Uniform Distribution Probability

$$P(X_1 \leq X \leq X_2) = \frac{X_2 - X_1}{b - a}$$

$$P(42 \leq X \leq 45) = \frac{45 - 42}{47 - 41} = \frac{1}{6}$$



Now, see the uniform distribution is defined in this interval 41 to 47 right. Suppose, we want to know the probability between 42 and 45, is it 42 and 45. So, that had not been done by $45 - x$, that is:

$$x_2 - x_1 / (b - a)$$

x_2 is 45, x_1 is 42 So,

$$45 - 42 = 3, 3 \text{ divided } 6 = 1/2. \text{ So, this area is } 0.5.$$

(Refer Slide Time: 20:04)

Example : Uniform Distribution

- Consider the random variable x representing the flight time of an airplane traveling from Delhi to Mumbai.
- Suppose the flight time can be any value in the interval from 120 minutes to 140 minutes.
- Because the random variable x can assume any value in that interval, x is a continuous rather than a discrete random variable

You will see another example of this uniform distribution. Consider a random variable x representing the flight time of an airplane travelling from Delhi to Mumbai. Suppose the flight

time can any value in the interval between 122 to 140 minutes. Because of the random variable x can assume any value in the interval, x is a continuous rather than a discrete random variable.

(Refer Slide Time: 20:30)

Example : Uniform Distribution contd....

- Let us assume that sufficient actual flight data are available to conclude that the probability of a flight time within any 1-minute interval is the same as the probability of a flight time within any other 1-minute interval contained in the larger interval from 120 to 140 minutes. 
- With every 1-minute interval being equally likely, the random variable x is said to have a uniform probability distribution.



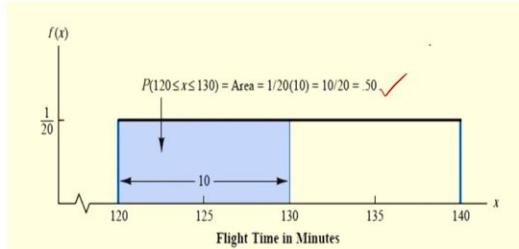
30

Let us assume that sufficient actual flight data are available to conclude that the probability of your flight time within any one minute interval is the same as the probability of a flight at time within any other one minute interval, contained in the large interval from 120 to 140. So, that is the properties of your uniform distribution. That means, in any one, any small interval you should have the same value.

If any one interval, if any between this interval the probability will say next, another one minute interval, the probability same. With every one minute interval being equally likely the random variable x is set to have a uniform probability distribution. So, the upper limit is 140, the lower limit is 120, $1/(140 - 120)$.

(Refer Slide Time: 21:20)

Probability of a flight time between 120 and 130 minutes



32

So, the height of this rectangle distribution is 1 divided by 20, it is starting a value is 120, b value is 140. Suppose, you are asked to find out, probability of your flight time between 120 and 130 minutes. So, somebody is asking what is the probability of that, flight arriving between 120 and 130 minutes. So, what you have to do, this is $(130 - 120)$ divided by $(140 - 120)$. When you simplify you will get 0.50. So the probability of that flight arrive between 120, 130 minutes is 50%.

(Refer Slide Time: 22:05)

Exponential Probability Distribution

- The exponential probability distribution is useful in describing the time it takes to complete a task.
- The exponential random variables can be used to describe:

Time between vehicle arrivals at a toll booth



Time required to complete a questionnaire



Distance between major defects in a highway



Next we will go to the exponential probability distribution. The exponential probability distribution is useful in describing the time it takes to complete a task. The exponential random variable can be used to describe the time between vehicle arrivals at a toll booth, time required to

complete a questionnaire, the distance between major defects in a highway. For example, whenever the time between arrival and time required to complete the questionnaire, whenever the word the between comes. Then that is the then it is appropriate to use exponential distribution.

(Refer Slide Time: 22:43)

Exponential Probability Distribution

- Density Function

$$f(x) = \frac{1}{\mu} e^{-x/\mu}$$

where: μ = mean
 $e = 2.71828$



The density function for a exponential probability distribution is $f(x) = (1/\mu) \cdot e^{(-x/\mu)}$. Here, μ is the mean.

(Refer Slide Time: 22:52)

Exponential Probability Distribution

- Suppose that x represents the loading time for a truck at loading dock and follows such a distribution.
- If the mean, or average, loading time is 15 minutes ($\mu = 15$), the appropriate probability density function for x is

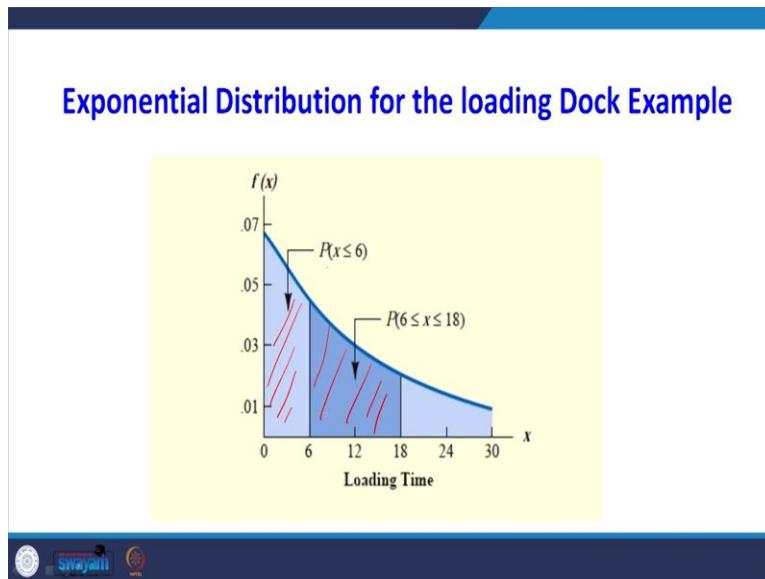
$$f(x) = \frac{1}{15} e^{-x/15}$$



We will see what is the how to construct the exponential distribution. Suppose that x represents the loading time for a truck at a loading dock and follow such a distribution. If the mean or

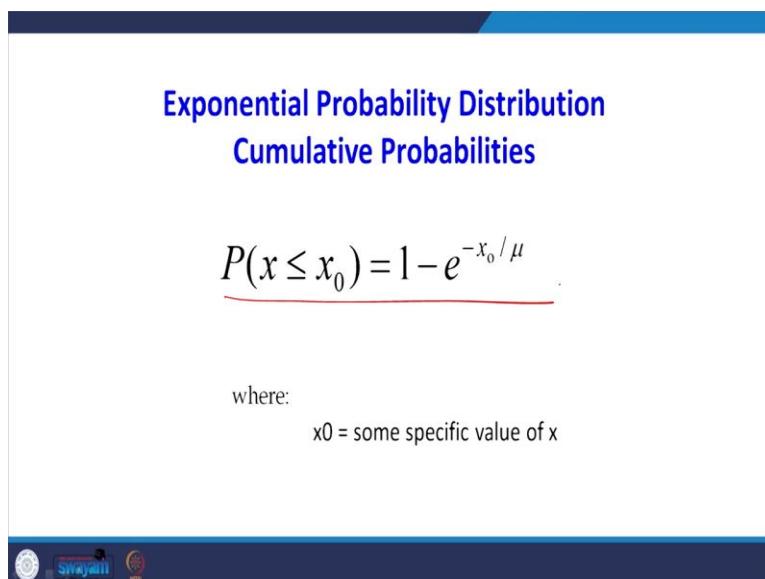
average loading time is 15 minutes, $\mu = 15$, then appropriate probability density function is f of x = 1 divided by 15 e to the power – x divided by 15.

(Refer Slide Time: 23:13)



So, what is the meaning of this exponential distribution is supposed, if the loading time, the probability of loading time less than 6 is this area. The probability of loading time between 6 and 18 is this one this time, this is a probability, this shadowed portion represents the probability of that loading time is between 6 and 18.

(Refer Slide Time: 22:39)



Because in many application of exponential distribution will use the cumulative probability density function. So, for an exponential distribution, the formula for finding the cumulative

probability density function = $1 - e^{(-x_0 / \mu)}$ where x_0 is some specific value of x . It is nothing but if you integrate that to distribution, if you integrate the distribution between the intervals when you simplify it, you will get this answer.

(Refer Slide Time: 24:08)

Example: Exponential Probability Distribution

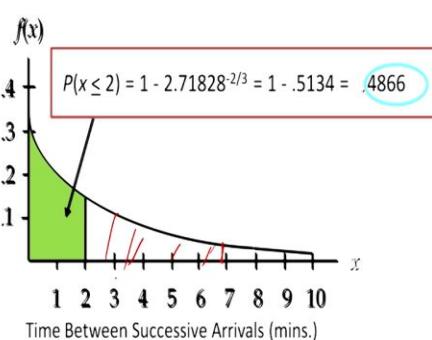
- The time between arrivals of cars at a Petrol pump follows an exponential probability distribution with a mean time between arrivals of 3 minutes.
- The Petrol pump owner would like to know the probability that the time between two successive arrivals will be 2 minutes or less.



We will see one example of this exponential probability distribution. The time between arrivals of cars at a petrol pump follows an exponential probability distribution with the mean time between arrivals of 3 minutes. See that it is mean time between arrivals. So, the mean is 3 here. The petrol pump owner would like to know, the probability that time between 2 successive arrivals will be 2 minutes or less. So x value is less than or equal to 2.

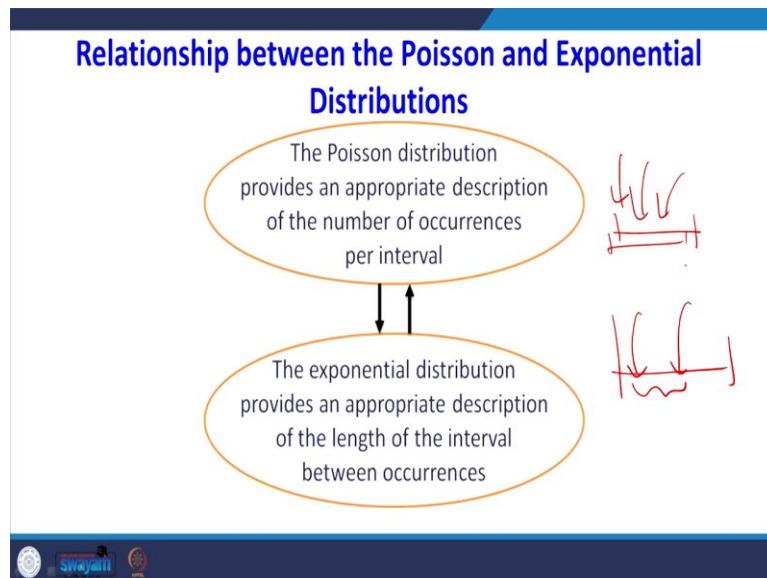
(Refer Slide Time: 24:43)

Example: Petrol Pump Problem



So if you want to know P of x less than equal 2, we have to substitute into the, that distribution. That is the cumulative to distribution function, $1 - x$ value, so it is 0.4866. It is a 0.4866 so this one. So this shaded one. Suppose if you want to find out, the time between 2 successive arrivals of vehicle is less than 7 minutes, probability will increase. So, when x increases, but there are more chances that the time between 2 success arrival is this much. So what I am saying this is a way to interpret the exponential distribution.

(Refer Slide Time: 25:23)



Now, there is a very important relationship between poisson and exponential distribution. See, the Poisson distribution provides an appropriate description of number of occurrences per interval. So, one interval is there. In that interval, how many occurrences happened, that is the Poisson distribution. In the exponential distribution provides an appropriate description of length of the interval between the occurrences.

Suppose the same interval between this occurrence to this occurrences, so this phenomena is explained with the help of between 2 occurrences, this phenomena is explained with the help of exponential distribution. Number of occurrences in that interval it is explained with the help of Poisson distribution, that is a relation between poisson and exponential distribution.

(Refer Slide Time: 26:14)

Mean of Poisson and Mean of Exponential Distributions

- Because the average number of arrivals is 10 cars per hour, the average time between cars arriving is

$$\frac{1 \text{ hour}}{10 \text{ cars}} = .1 \text{ hour/car}$$

10
 $\mu \rightarrow$ mean for poiss
 $\lambda \rightarrow$ mean for exp
 μ

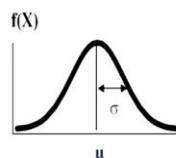


So, mean of your poisson and mean of your exponential distribution, what is the relationship. Because the average number of arrival is 10 cars per hour, the average. So, if it is a 10 in that interval, there are 10 cars are arriving. So, this mean is taken for the Poisson distribution. But 1 / 10 is taken as the mean for the exponential distribution. That means, time between arrivals. So, that is a link between μ generally we write μ mean for Poisson distribution, $1 / \mu$ mean for exponential distribution that is the relation between poisson and exponential distribution.

(Refer Slide Time: 27:05)

The Normal Distribution: Properties

- 'Bell Shaped'
- Symmetrical
- Mean, Median and Mode are equal
- Location is characterized by the mean, μ
- Spread is characterized by the standard deviation, σ
- The random variable has an infinite theoretical range: $-\infty$ to $+\infty$



Mean = Median = Mode



Next, we are entering into the very important distribution that is a normal distribution, normal distribution is we can say a father of all the distributions. Because suppose some phenomena is happening if you are not aware, you can assume that it follows normal distribution. Normal

distribution is following Bell Shaped, it is symmetrical, mean, median and modes are equal. The location of the normal distribution is characterized by μ , the spread is characterised by σ . The random variable has an infinite theoretical range that is minus infinity to plus infinity.

(Refer Slide Time: 27:50)

The Normal Distribution: Density Function

The formula for the normal probability density function is

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{(X-\mu)}{\sigma}\right)^2}$$

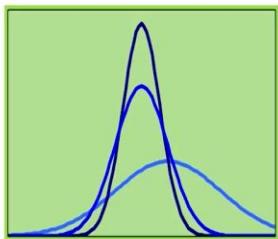
Where e = the mathematical constant approximated by 2.71828
 π = the mathematical constant approximated by 3.14159
 μ = the population mean
 σ = the population standard deviation
 X = any value of the continuous variable

Chap 6-43

The density function of a normal distribution is $f(X) = (1/\sqrt{2\pi}\sigma) e^{-(x-\mu)^2/(2\sigma^2)}$, where e is the mathematical constant value is 2.71, π is the mathematical constant we know that 3.14, μ is the population mean, σ is the population standard deviation, X is any value of the continuous variable.

(Refer Slide Time: 28:18)

The Normal Distribution: Shape



By varying the parameters μ and σ , we obtain different normal distributions

Chap 6-43

The shape of the normal distribution is by varying the parameter of μ and σ we obtained different normal distributions. Dear students, what we have seen so far is we have seen some of the discrete and continuous distributions in the discrete distribution we have talked about the binomial, Poisson distribution. In the continuous distribution we have seen exponential and uniform distributions. The next class very important distribution that is normal distribution, that will cover in the next class. Thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology Roorkee

Lecture – 10
Probability Distributions - III

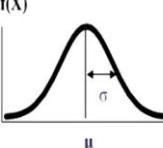
Welcome back students now we are going to discuss another important continuous distribution and that is normal distribution, normal distribution can be called as mother of all distribution because, if in any phenomena if you are not aware about the nature of the distributions, you can assume that it follows normal distribution, most of the statistical test or whatever analytical tools which are going to use in this course.

Good to have some assumptions that did follow normal distributions knowing the properties and behavior and assumptions about the normal distribution is very important for this course. Some of the properties: normal distribution is Bell shaped curved.

(Refer Slide Time: 01:16)

The Normal Distribution: Properties

- 'Bell Shaped'
- Symmetrical
- Mean, Median and Mode are equal
- Location is characterized by the mean, μ
- Spread is characterized by the standard deviation, σ
- The random variable has an infinite theoretical range: $-\infty$ to $+\infty$



Mean = Median = Mode

Right we form a bell shaped curve. It is symmetrical, you can fold it so after folding both the sides are same, another important property mean, median and modes are equal the location is characterized by its mean μ the spread is characterized by standard deviation. The random variable has an infinite theoretical range that is minus infinity to plus infinity.

(Refer Slide Time: 01:48)

The Normal Distribution: Density Function

The formula for the normal probability density function is

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{(X-\mu)}{\sigma}\right)^2}$$

Where e = the mathematical constant approximated by 2.71828

π = the mathematical constant approximated by 3.14159

μ = the population mean

σ = the population standard deviation

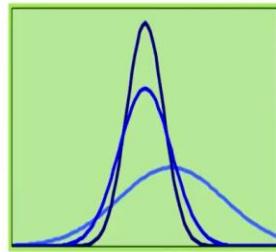
X = any value of the continuous variable



The formula for normal probability density function $f(X) = (1/\sqrt{2\pi\sigma^2}) e^{-(x-\mu)^2/(2\sigma^2)}$, where e is the mathematical constant, the value is 2.71828 π is the mathematical constant the value is 3.14 μ is the population mean σ is the population standard deviation, X is any value of the continuous variable.

(Refer Slide Time: 02:18)

The Normal Distribution: Shape



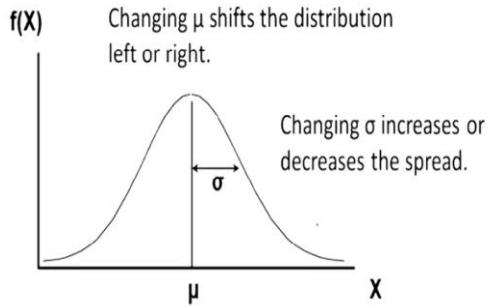
By varying the parameters μ and σ , we obtain different normal distributions



The shape of the normal distribution will change based on its spread by varying parameters μ and σ being obtain different normal distributions. For example this one where the sigma is very low, in this case sigma is little normal, in this is sigma is very big.

(Refer Slide Time: 02:39)

The Normal Distribution: Shape



Changing μ shift to the distribution left or right if you increase the value of μ , it can go right side or left side. Changing σ , the standard deviation increases or decreases the spread generally, when you decrease the σ the spread will decrease when you increase the σ the spread will increase.

(Refer Slide Time: 03:01)

The Standardized Normal Distribution

- Any normal distribution (with any mean and standard deviation combination) can be transformed into the standardized normal distribution (Z).
- Need to transform X units into Z units.
- The standardized normal distribution has a mean of 0 and a standard deviation of 1.



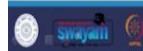
There is another normal distribution standardized normal distribution, any normal distribution with the mean and standard deviation combination can be transformed into standardized normal distribution. One thing what you had to do, we need to transform X unit into Z units, Z is nothing but the conversion method is $(X - \mu) / \sigma$, the standardized normal distribution as means 0 and the variance or standard deviation is 1.

(Refer Slide Time: 03:36)

The Standardized Normal Distribution

- Translate from X to the standardized normal (the "Z" distribution) by subtracting the mean of X and dividing by its standard deviation:

$$Z = \frac{X - \mu}{\sigma}$$



The translation from X to the standardized normal that is the Z distribution by subtracting the mean of the X and dividing by standard deviation. So, that conversion from it is normal distribution to standardized normal distribution is done with the help of this Z transformation
Where $Z = (X - \mu) / \sigma$

X is a random variable mu (μ) is the mean of the population. Sigma(σ) is the standard deviation of the population.

(Refer Slide Time: 04:05)

The Standardized Normal Distribution: Density Function

- The formula for the standardized normal probability density function is

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}}$$

Where e = the mathematical constant approximated by 2.71828

π = the mathematical constant approximated by 3.14159

Z = any value of the standardized normal distribution



The formula for the standardized normal probability density function, if you substitute Z equal to $X - \mu / \sigma$ in our previous equation and it become if

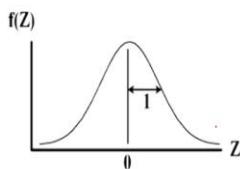
$$f(Z) = \left(1/\sqrt{2\pi}\right) e^{(-Z^2)/2}$$

Where π is the mathematical constant z is any value of this standardized normal distribution.

(Refer Slide Time: 04:26)

The Standardized Normal Distribution: Shape

- Also known as the "Z" distribution
- Mean is 0
- Standard Deviation is 1



Values above the mean have positive Z-values, values below the mean have negative Z-values



Standardized normal distribution the shape how they look like also known as Z distribution mean is 0 standard deviation is 1, the value above the mean has positive Z value, values below the mean will have negative Z value.

(Refer Slide Time: 04:44)

The Standardized Normal Distribution: Example

- If X is distributed normally with mean of 100 and standard deviation of 50, the Z value for $X = 200$ is

$$Z = \frac{X - \mu}{\sigma} = \frac{200 - 100}{50} = 2.0$$

- This says that $X = 200$ is two standard deviations (2 increments of 50 units) above the mean of 100.

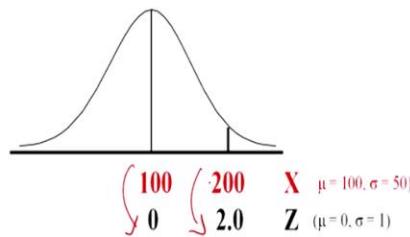


Let us see how to do that conversion from normal distribution to standardized normal distribution. If X is distributed normally with the mean of 100 and standard deviation of 50 the Z value of X is 200 then corresponding Z value is $X - \mu / \sigma$, X is 200, $- \mu$ 100, divided by

sigma 50, equal to 2.0. This says that $X = 200$ is two standard deviation above the mean of 100, that is 2 increments of 50 units, the Z value nothing but how many times of it is standard deviation that is nothing but your Z, here 2 increments of 50 that is why the Z value is 2.

(Refer Slide Time: 05:41)

The Standardized Normal Distribution: Example



Note that the distribution is the same, only the scale has changed. We can express the problem in original units (X) or in standardized units (Z).



Look at the conversion now this will be so convenient for you the red one, where the mean = 0 the $X = 200$, we have asked to find out when $X = 200$ what is the corresponding Z value? The red will shows in the simple normal distribution, the black one shows the standardized normal distribution, you see that the mean of the distribution is under mu standardized scale it becomes 0, when $X = 200$ in a normal distribution in a standardized normal distribution.

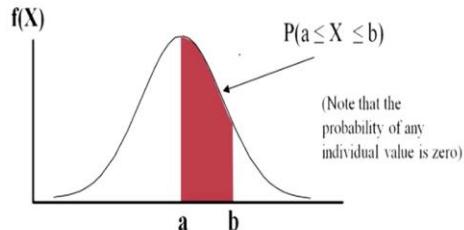
The X and corresponding Z value is 2 , where the mean mu equal to 0 sigma equal 1. Note that the distribution is the same only the scale is has changed. We can express the problem in original units are standardized units but there is an advantage. Why we have to convert into standardized normal distribution sometime you may be required to find out the area of a distributions. Because if you are not standardizing you cannot use that your Z table, Z statistical table. Every time to know the area you have to integrate.

That is a very compression process that is why every normal distribution is converted to standardized normal distribution for the convenient of looking at the Z value directly from the table that will simplify our task.

(Refer Slide Time: 07:13)

Normal Probabilities

Probability is measured by the area under the curve

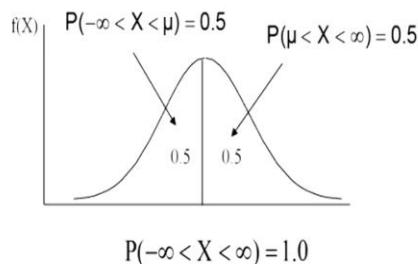


The probability is measured by area under the curve in a continuous distribution, the probability you know that it is measured area under the curve suppose, always it has to be expressed between A and B. If you want to know the probability exactly at A are exactly B that will not form the area. So, the probability is 0. So in the context of continuous distribution, the meaning of probabilities area under the curve, but if it is a discrete probability distributions, the probability can be read directly by looking at the X and corresponding P of X.

(Refer Slide Time: 07:53)

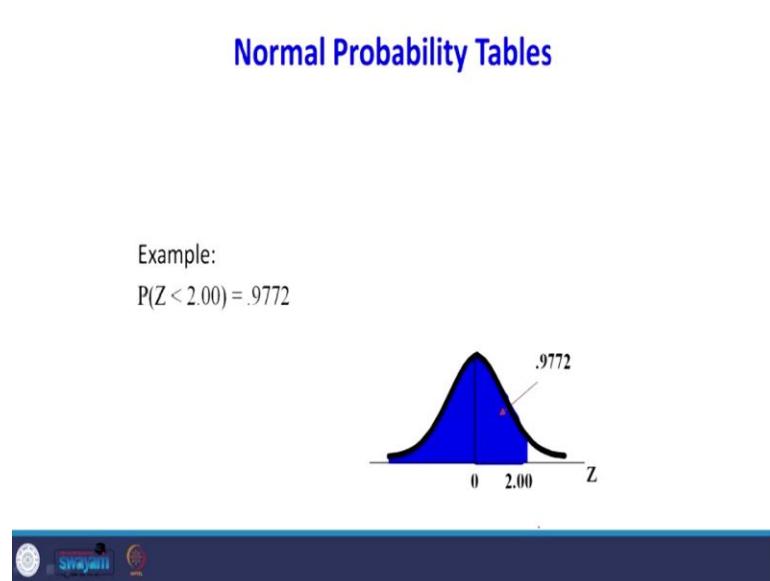
Normal Probabilities

The total area under the curve is 1.0, and the curve is symmetric, so half is above the mean, half is below.



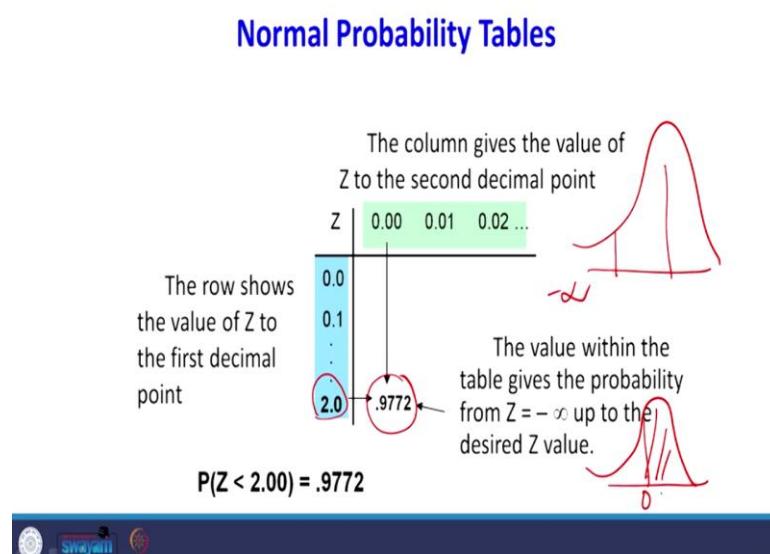
Total area under the curve is 1 and the curve is symmetric, so half is above the mean half is below. So $P(-\infty \leq X \leq \mu)$ is 0.5 similarly, $P(\mu \leq X \leq \infty)$ is 0.5, so the total area is 1.

(Refer Slide Time: 08:16)



Suppose, if you want to know the area Z less than 2.00, see this was when the Z is lesser, this area is 0.9772.

(Refer Slide Time: 08:30)



One way you can read it directly from the Z table suppose in the rows, the Z value is given the column the decimal of it is given. Suppose if you want to know $Z = 2.00$ you have to look at in row 2.00, the corresponding area is this one. See, the rows shows the value of Z to the first

decimal point. The column gives the value of Z to the second decimal point. The value within the table gives the probability from Z minus infinity up to desired Z value.

When we look at the table, statistical table especially Z table it should be very careful whether the area is given from minus infinity, there are 2 possibilities sometime the area may be given minus infinity to plus X value, sometime area may be given only the positive value, this side value 0, positively values of Z is given. If you want to know if you want to read the negative value of Z because it is symmetric, so you all can read just only the positive value, then we can take that value to the negative side.

(Refer Slide Time: 09:53)

**Finding Normal Probability
Procedure**

To find $P(a < X < b)$ when X is distributed normally:

- Draw the normal curve for the problem in terms of X.
- Translate X-values to Z-values.
- Use the Standardized Normal Table.

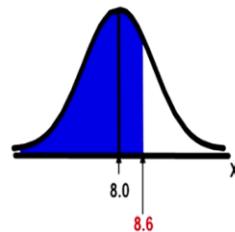


So, finding normal probability procedure we will see one problem to find $P(a < X < b)$ when X is distributed normally, the first one is draw the normal curve of the problem in terms of X, whenever you are going to find out area, it is always good to draw the distribution draw the normal distribution then you can intuitively you can read from the picture, so the next step is translate X value to Z values then use standardized normal table where you can get the area.

(Refer Slide Time: 10:34)

Finding Normal Probability: Example

- Let X represent the time it takes (in seconds) to download an image file from the internet.
- Suppose X is normal with mean 8.0 and standard deviation 5.0
- Find $P(X < 8.6)$



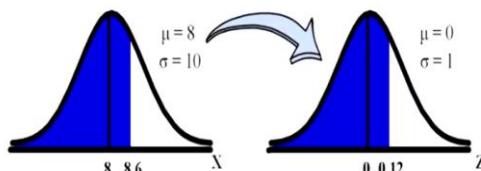
Let X represents the time it takes to download an image file from the internet. Suppose X is a normal with mean 8 and standard deviation 5. If we want to know what is the probability of X less than 8.6 that means what is the probability of downloading time is below 8.6 So first you have to mark the mean then you are to find out this X values 8.6, so since it is asked less than 8.6, the left side area, so the first steps is 8.6 has to be converted into, you can integrated by using normal distributions, you can substitute to minus infinity to 8.6 mean you can integrated, we will get the area there is no problem, but it is very time consuming process.

(Refer Slide Time: 11:26)

Finding Normal Probability: Example

- Suppose X is normal with mean 8.0 and standard deviation 5.0. Find $P(X < 8.6)$.

$$Z = \frac{X - \mu}{\sigma} = \frac{8.6 - 8.0}{5.0} = 0.12$$



So, one easy way is you have to convert that normal distribution into standard normal distribution, that means the X value has to be converted into Z scale they can read they can use

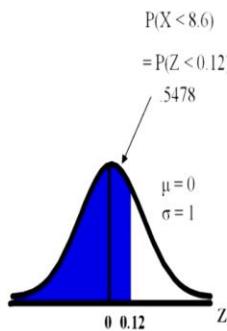
the table to find out the area for the corresponding Z value. Suppose, X is the normal with mean 8 or standard deviation 5. X less than 8.6 use the $Z = (X - \mu)/\sigma$, formula to get Z value when $X = 8.6$. So we got 0.12, so now when Z value 0.12 you can read this value directly from the normal table to know the probability.

(Refer Slide Time: 12:06)

Finding Normal Probability: Example

Standardized Normal Probability
Table (Portion)

Z	.00	.01	.02
0.0	.5000	.5040	.5080
0.1	.5398	.5438	.5478
0.2	.5793	.5832	.5871
0.3	.6179	.6217	.6255



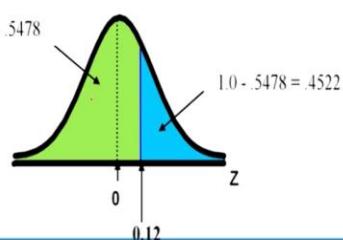
Is it that Z value 0.12 so you can Z value 0.12 so this area is 0.5748.

(Refer Slide Time: 12:15)

Finding Normal Probability: Example

- Find $P(X > 8.6)$...

$$\begin{aligned} P(X > 8.6) &= P(Z > 0.12) = 1.0 - P(Z \leq 0.12) \\ &= 1.0 - .5478 = .4522 \end{aligned}$$



Finding normal probability suppose X is greater than 8.6, so now we have to look at the area of the right side so, P of X greater than 8.6 is equal to that we have to convert it to Z scale after getting since it is greater than since the area is 1.

$1 - P(Z \text{ less than } 0.12)$, will give the blue side area. So, one when $Z = 0.12$ corresponding areas 0.54 so, this side area is after subtracted from one will get, we are getting 0.4522.

(Refer Slide Time: 12:56)

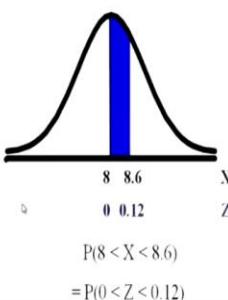
Finding Normal Probability: Between Two Values

- Suppose X is normal with mean 8.0 and standard deviation 5.0.
Find $P(8 < X < 8.6)$

Calculate Z-values:

$$Z = \frac{X - \mu}{\sigma} = \frac{8 - 8}{5} = 0$$

$$Z = \frac{X - \mu}{\sigma} = \frac{8.6 - 8}{5} = 0.12$$



$$= P(0 < Z < 0.12)$$



Suppose X is a normal with mean 8 standard deviation 5 so find $P(8 \text{ less than } X \text{ less than } 8.6)$ now, the 2 value of X is given both of values has to convert when $X = 8$ we are getting Z value 0, when $X = 8.6$ we are getting Z value 0.12, so now we have to know the area of $Z = 0$, to $Z = 0.12$. So that means 0 to 0.12.

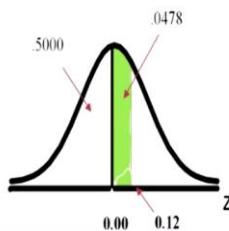
(Refer Slide Time: 13:28)

Finding Normal Probability Between Two Values

- Standardized Normal Probability
- Table (Portion)

Z	.00	.01	.02
0.0	.5000	.5040	.5080
0.1	.5398	.5438	.5478
0.2	.5793	.5832	.5871
0.3	.6179	.6217	.6255

$$\begin{aligned} P(8 < X < 8.6) \\ &= P(0 < Z < 0.12) \\ &= P(Z < 0.12) - P(Z \leq 0) \\ &= .5478 - .5000 = .0478 \end{aligned}$$

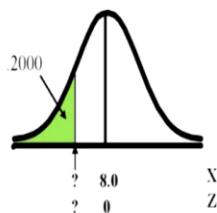


One way from the table is, first you find the area up to minus infinity to Z value 0.12. So, we are getting 0.5478 then subtract when $Z = 0$ left side area we know it is a .5. So, the remaining is 0.0478.

(Refer Slide Time: 13:50)

Given Normal Probability: Find the X Value

- Let X represent the time it takes (in seconds) to download an image file from the internet.
- Suppose X is normal with mean 8.0 and standard deviation 5.0
- Find X such that 20% of download times are less than X .



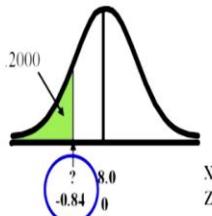
Now, just the reverse of that the probability is given you have to find out the X value, let X represents the time It takes to download an image file from the internet suppose X is normal with mean 8 standard deviation 5 find X such that 20% of the download times are less than X , there are 2 points here, one is less than X another one is 20%. So, on the left hand side when area equal to 0.2 what is the corresponding X value so, for that first you got to find out Z value, from the Z you have to find out the X value.

(Refer Slide Time: 14:35)

Given Normal Probability, Find the X Value

- First, find the Z value corresponds to the known probability using the table.

Z03	.04	.05
-0.91762	.1736	.1711
-0.82033	.2005	.1977
-0.72327	.2296	.2266

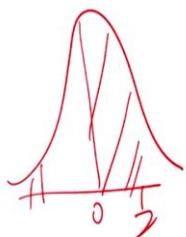


Now look at the table. So, when area equal to 0.2 corresponding Z value is (-0.84), this is the value of Z.

(Refer Slide Time: 14:45)

Given Normal Probability, Find the X Value

- Second, convert the Z value to X units using the following formula.



$$\begin{aligned}
 X &= \mu + Z\sigma \\
 &= 8.0 + (-0.84)5.0 \\
 &= 3.80
 \end{aligned}
 \quad \text{Z} = \frac{X - \mu}{\sigma}$$

So 20% of the download times from the distribution with mean 8.0 and standard deviation 5.0 are less than 3.80 seconds.



So, we know that Z value is (-0.84), here this formula has come from this simple formula = $(X - \mu) / \sigma$. Now, we know the value of Z from this you have to find out value of X. And one more thing the when you are finding the value of Z, you should be very careful what kind of normal distribution you are using to find out the value of Z if normal distribution is like this that is area is given from 0 to positive Z right. So if you are measuring area on the left hand side, so will get them Z value but have to attach negative side to that. So we should be careful.

So $\mu = 8.0$,

$X = 8.0 + (-0.84)5$, we are getting 3.80. So 20% of the download times from the distribution with the mean 8 and standard deviation 5 are less than 3.8 seconds.

(Refer Slide Time: 15:47)

Assessing Normality

- It is important to evaluate how well the data set is approximated by a normal distribution.
- Normally distributed data should approximate the theoretical normal distribution:
 - The normal distribution is bell shaped (symmetrical) where the mean is equal to the median.
 - The empirical rule applies to the normal distribution.
 - The interquartile range of a normal distribution is 1.33 standard deviations.



Another important thing gives us is normality because the normality assumption is very important for other type of inferential statistics. I will tell you why it is important because we will be studying a concept called Central Limit Theorem, where when you do the sampling of the sampling that will follow normal distributions. So, lot of many analytical tools many statistical tools follow the assumption that data should follow normal distributions, that is why as soon as you collect the data.

The first step is cleaning the data, when the cleaning in that process is you have to verify whether the data follow normal distribution or not, otherwise, you may not otherwise you will you may end up choosing wrong statistical techniques or analytical techniques.

(Refer Slide Time: 16:35)

Assessing Normality

- It is important to evaluate how well the data set is approximated by a normal distribution.
- Normally distributed data should approximate the theoretical normal distribution:
 - The normal distribution is bell shaped (symmetrical) where the mean is equal to the median.
 - The empirical rule applies to the normal distribution.
 - The interquartile range of a normal distribution is 1.33σ .



It is important to evaluate how well the data Z is approximated by a normal distribution. Normally distributed data should approximate theoretical normal distribution, like the normal distribution is bell shaped where the mean is equal to the median. The empirical rule applies to the normal distribution. The interquartile range of a normal distribution is 1.33σ ; these are the way to test the normality.

(Refer Slide Time: 17:04)

Assessing Normality

- Construct charts or graphs
 - For small- or moderate-sized data sets, do stem-and-leaf display and box-and-whisker plot look symmetric?
 - For large data sets, does the histogram or polygon appear bell-shaped?
- Compute descriptive summary measures
 - Do the mean, median and mode have similar values?
 - Is the interquartile range approximately 1.33σ ?
 - Is the range approximately 6σ ?



$S \approx 0$



Another way to assess the normality is construct the charts or graph. Now, you can look at the shape of the distribution, for small or moderate sized data set, do stem and leaf display and box and whisker plot and check whether it is look symmetrical or not. As I told you in the beginning

of the lectures, if you look at the stem and leaf plot, you should follow this kind of shape then we can say it follows normal distributions, In the box and whisker plot.

For example, box and whisker plot is like this, right, the middle line that is median line should be in the middle of the box then only we can say the data follows normal distribution for a large data set. That is the histogram or polygon appears bell shaped, you can draw a histogram and also you can verify whether it follows normal distribution. Other way you can compute descriptive summary measures, whether you can check mean median mode.

How the similar value is the interquartile range approximately 1.33 sigma is the range is approximately 6 sigma, these are some descriptive measures to check whether the data follow normal distribution or not then you can find the skewness, when the skewness is 0 then we can say this data follow normal distribution.

(Refer Slide Time: 18:31)

Assessing Normality

- Observe the distribution of the data set
 - Do approximately 2/3 of the observations lie within $\text{mean} \pm 1$ standard deviation?
 - Do approximately 80% of the observations lie within $\text{mean} \pm 1.28$ standard deviations?
 - Do approximately 95% of the observations lie within $\text{mean} \pm 2$ standard deviations?



Some more examples, to check the normality observed the distribution of the data set these are the conditions do approximately 2/3 of the observations lie within the ± 1 sigma. Then we can see it follows normal distribution do approximately 80% of the observations lie within ± 1.28 standard deviations are do approximately 95% of the observations lie within the mean or ± 2 standard deviations, this is the Z table.

(Refer Slide Time: 19:02)

Z Table



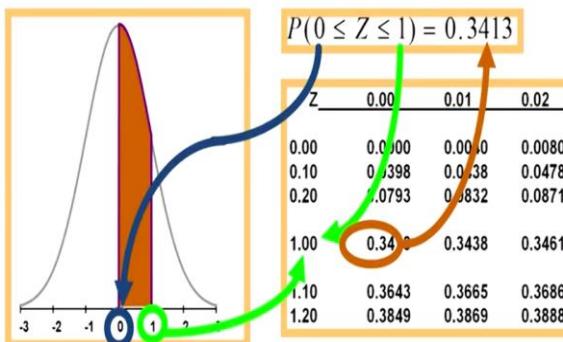
Second Decimal Place in Z											
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	
0.00	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359	
0.10	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753	
0.20	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141	
0.30	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517	
0.40	0.1559	0.1586	0.1612	0.1638	0.1664	0.1689	0.1715	0.1740	0.1765	0.1789	
1.00	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621	
1.10	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830	
1.20	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015	
2.00	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817	
3.00	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990	
3.40	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998	
3.50	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	



You see the previously the Z table is starting from 0 is not starting from minus infinity. So, this is second decimal, suppose Z is 0, the probability is 0, here it is given, what is given only one side known. The area is given only this one. So if you are finding you have to add 0.5 suppose if you want 0.0 you have to add 0.5 to get the, Z table. Another important .which I am planning to, willing to share with you.

(Refer Slide Time: 19:39)

Table Lookup of a Standard Normal Probability



See this $Z = 0$, $Z = 1$ see, this is 0.3413, right between 0 and one. Suppose if we want to know minus infinity to 1, you have to add 0.5 with that. Plus 0.5. So, we will get the value, another one when you see, when you look at the normal distribution. I will come back to that.

(Refer Slide Time: 20:04)

Applying the Z Formula

X is normally distributed with $\mu = 485$, and $\sigma = 105$

$$P(485 \leq X \leq 600) = P(0 \leq Z \leq 1.10) = .3643$$

For $X = 485$,

$$Z = \frac{X - \mu}{\sigma} = \frac{485 - 485}{105} = 0$$

For $X = 600$,

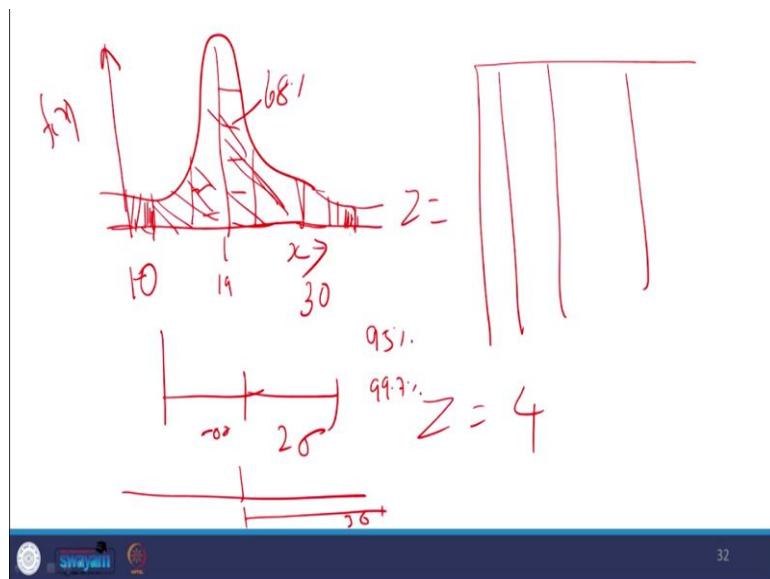
$$Z = \frac{X - \mu}{\sigma} = \frac{600 - 485}{105} = 1.10$$

Z	0.00	.01	.02
0.0	0.0000	0.0040	0.0080
0.1	0.0398	0.0438	0.0478
1.0	0.3413	0.3438	0.3461
1.10	0.3643	0.3665	0.3686
1.20	0.3849	0.3869	0.3888



If X is normally distributed with the mu = 485 and sigma = 105. So, the 485 to 600 when X is 485, you have to convert that scale to 0 when X = 600 corresponding Z values 1.1 so, Z is 0 to 1.1, 0.3643 is the area and the curve. Dear students we have seen So, far the properties of normal distribution then we have seen standardized normal distribution a normal distribution how these are interrelated and we have seen how to find out the area with the help of table one property.

(Refer Slide Time: 20:51)



You can look at the normal distribution the normal distribution shape is like this. So, you look at this, it would not touch this is x axis y is your probability effects. It will not touch the x axis you may have this doubt why it is not touching this distribution normal distribution why it is not touching axis? Because suppose, if I am plotting age of the students in the class follow normal

distribution, see the average age is say 19. There is a possibility somebody, suppose I am closing this way, some bodies age may be say around 30, somebody age may be around say, around 10.

So, since this normal distribution was drawn with the help of sample, I was not exactly knowing that this kind of rare value of X whether it is $X = 30$ hours $X = 10$. So, why I am not closing? Why this normal distribution not touching X axis, because we were given provision for the rare events that means X is maybe very high value X may be very low value, but I am not sure about that. That is why the normal distribution did not touch with the X axis. The another doubt you may know when you look at the Z table.

When you look at the Z table, the value of Z most of the time I go back. It will show you, see this the value Z is 3.5. The question may come why the value of Z is maximum 4 or 5 in the statistical table, you remember the beginning of the class, I was saying from the mean if you travel on either side with one sigma distance you can capture 68%, if you travel 2 sigma distance from the mean, that is this distance, 2 sigma distance and minus 2 sigma distances.

You can capture 95% of the area of the normal distribution. If you travel 3 sigma distances this extreme distance, I can use some other colour, please bear with me. If you travel 3 sigma distance, this portion, if you travel 3 sigma distance, you can cover 99.7% of the data. Okay. So, why the value of Z is not beyond 3, the possibility of the Z value is beyond 3 is only 0.3% the same time the probability of x value to become extremely high or extremely low is only 0.3%.

What is the meaning of that only 0.3% chance that the value of that will be more than 3, that is why all statistical tables given only 3.5 or 4 not beyond that. The another reason and also why we are not closing with the x axis the probability of that extreme events to happen is only 0.3% provided if it is following normal distribution. Now, I will summarize the students that so far via we have seen different type of probability distributions.

The previous class we have seen some of the continuous distribution in this class we have seen an important normal distribution that is a normal distribution. We have learned properties, normal distribution and a standard normal distribution, how to convert a normal distribution to

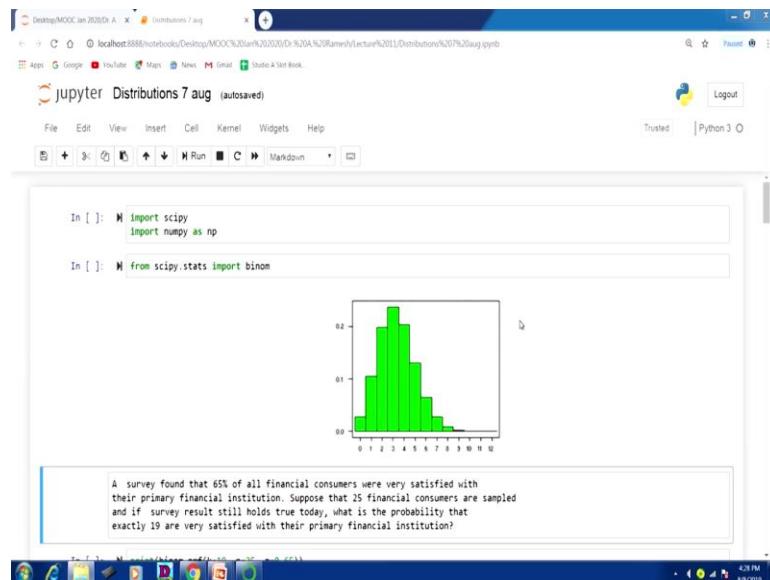
standard normal distribution, how to refer Z table to find out the area that you have seen. The next class with the help of Python will use how to find out the area under the curve or how to find out the mean standard deviation of different distributions. Thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 11
Python Demo for Distribution

Dear students in the last lecture we have seen probability distributions in this lecture with the help of Python we solve some problems from probability distributions.

(Refer Slide Time: 00:44)



The problem is taken from a book written by Ken Black the title of the book is applied statistics so we will see the problems now. Now I am importing scipy importing numpy as np from scipy dot stats import binom, binom is for doing binomial operations and one more thing you can import picture also for example this picture that is that empirical distribution picture I have taken from this source that is exclamation symbol square bracket that link and that link should not be in square bracket then you will when you do that link we can get that picture directly I am executing this.

Yes now we will see the problem a survey found that 65% of all financial consumers were very satisfied with their primary financial institution. Suppose that 25 financial consumers are sampled and if the survey result still hold the true today what is the probability that exactly 90 are very satisfied with their primary finance institutions. By looking at the problem we have to

see what kind of distribution we are going to use here there are 2 possibilities satisfied or not satisfied.

So, there are 2 possibilities there so then we can go for binomial distributions print binom dot pmf, probability mass function k equal to 19 that is we say in the probability distribution current context, n equal to number of sample p is probability of success. Now we can see that the answer is 0.09 so there is 0.09 the probability that exactly nine are satisfied with a primary financial institutions.

(Refer Slide Time: 02:43)

The screenshot shows a Jupyter Notebook interface with the title 'jupyter Distributions 7 aug (unsaved changes)'. The notebook contains the following content:

```
In [3]: M print(binom.pmf(k=19, n=25, p=0.65))
Out[3]: 0.0907779985932791
```

According to the U.S. Census Bureau, approximately 6% of all workers in Jackson, Mississippi, are unemployed. In conducting a random telephone survey in Jackson, what is the probability of getting two or fewer unemployed workers in a sample of 20?

```
In [4]: M binom.cdf(2, 20, 0.06)
Out[4]: 0.8850275957378545
```

Solve the binomial probability for n = 20, p = 0.4, and x = 10

```
In [ ]: M print(binom.pmf(k=10, n=20, p=0.4))
Out[ ]: 0.158462107439
```

Poisson Distribution

We go to the next problem this Book this problem also taken from that book, according to US Census Bureau approximately 6 percentage of all workers in Jackson Mississippi are unemployed in conducting a random telephone survey in Jackson what is the probability of getting 2 or fewer unemployed workers in a sample of 20. Here we want to know 2 are less so say the probability of 0 plus probability of 1 plus probability of 2.

So we were to do the cumulative density function. So, here for doing that one you have to enter type binom.cdf(2, 20, 0.6) the 2 represents less than or equal to 2, 20 represents the sample, the p represents the probability. So, when you run this you are getting community probability of 88.5%.

We will take another problem solve the binomial probability n equal to 20 sample size is 20 p equal to 0.4 x equal to 10 so binom dot pmf you will get the answer for 0.117.

We well go to the next distribution Poisson distribution. So, in the Poisson distribution for doing Poisson distribution you have to import the library Poisson distribution from scipy dot stats import Poisson first we will find out Poisson probability mass function so poisson.pmf (3, 2). 3 represents the x and 2 represent the mean. We will see another problem, suppose bank customers arrives randomly on any weekday afternoon at an average of 3.2 customers every 4 minutes, what is the probability of exactly 5 customers arriving in a 4 minute interval on a weekday afternoon.

By looking at the problem you say that we know that the arrival pattern follow Poisson distribution and you have to be very careful on the unit of mean and the unit of x both are in 4 minutes then no problems they simply can poisson.pmf(5, 3.5) is your x value Poisson dot pmf so 3.2 is the arrival rate, so 5,3 that is the 11.39%. You will see one more problem, bank customers arrive randomly on weekday afternoon at an average of 3.2 customers every 4 minutes what is the probability of having more than 7 customers in you 4 minute interval on a week day afternoon.

So, here we have to find out the probability of x greater than 7 so what we will do first we will find up to 7 with the help of this world that we will save any object called prob, equal to poisson dot cdf 7 and lambda 3.2 so when you subtract 1 minus of this 1 then we will get probability of more than that, yes so I am finding up to 7, when you substrate 1 minus that up to 7 we will get to more than 7. we will see another problem on Poisson.

On a bank has an average random arrival rate of 3.2 customers every 4 minutes what is the probability of getting exactly 10 customers during 8 minutes interval. now it should be very careful here the unit of x and unit of lambda are different, because it's a 4 minutes it is 8 minutes so you have to convert into same units so multiply by 32 by 2 you will get 6.4, so lambda equal to 10 so Poisson dot pmf of 10, 6.4 will give you the answer for 0.0527.

(Refer Slide Time: 06:47)

```

In [13]: poisson.pmf(18,6.4)
Out[13]: 0.052790043854115495

Uniform Distribution

Suppose the amount of time it takes to assemble a plastic module ranges from 27 to 39 seconds and that assembly times are uniformly distributed. Describe the distribution. What is the probability that a given assembly will take between 30 and 35 seconds?

In [ ]: U= np.arange(27, 39, 1)
U
In [ ]: from scipy.stats import uniform
uniform.mean(loc=27,scale=12)
In [ ]: uniform.cdf(np.arange(30, 35, 1), loc=27, scale=12)
In [ ]: Prob = 0.666666667 - 0.25
Prob

```

According to the National Association of Insurance Commissioners, the average annual cost for automobile insurance in the United States in a recent year was \$691. Suppose automobile insurance costs are uniformly distributed in the United States with a range of from \$200 to \$1,182. What is the standard deviation?

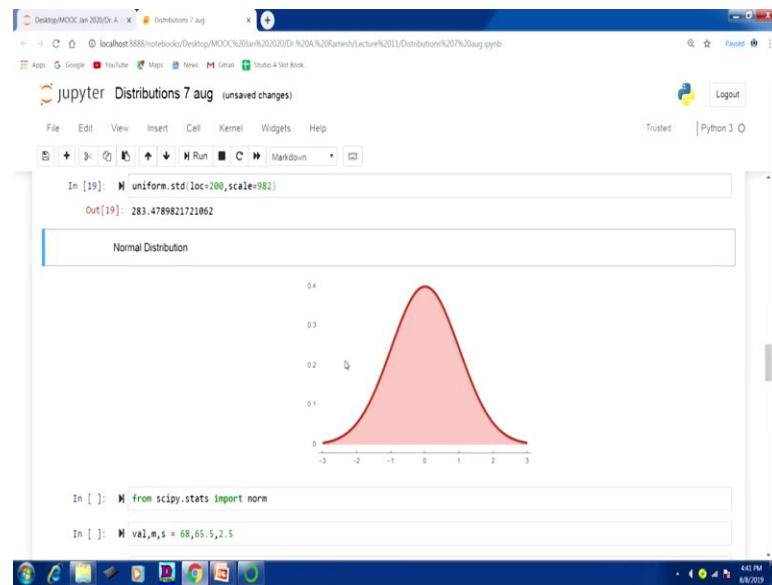
We will go to uniform distribution next suppose the amount of time it takes you see that the amount of time it takes to assembly a plastic module ranges from 27 to 39 seconds and the assembly time are uniformly distributed describe the distribution what is the probability that a given assembly will take between 30 to 35. first we will develop that array, so here u equal to np dot arrange, there are 2 functions one is range another one is arrange, arrange means its array, array function if you type if you type simply range that is a list.

So 27 is the starting value say '40-1' because it is $n - 1$, not 40, 39 will be the last value in the array, the increment is by 1, we got this one, now this is our uniform distribution, now from scipy dot start import uniform so we will find out the mean of this distribution so for that purpose uniform dot mean loc is the starting point 27, scale is how much plus 12 so it is a $27 + 12$ is 39. So, that is a syntax for, so the mean is 33 otherwise in the uniform distribution finding the mean is not a very complicated formula simply you have to find out the $(a + b) / 2$.

Then we will do the cumulative distribution cdf, cumulative function, so uniform dot cdf np dot a arrange 30 what the question was asked is 30 to 35. So, np dot array 30 because it is a 35 you have to go out to 36 the increment is 1, starting point is 27 this scale is 12. so this will give you the probability between 30 to 35 so, when you run this so the probability of 30 is 0.25, 31 is 0.33, 32 is 0.41 and so on. Suppose we want between 30 and 35 for 30 the probability is 0.25 for 35 it

is 0.66, so if you substrate 0.66 - 0.25 you will get the; and so far that probability that the given assembly will take between 30 to 35 seconds.

(Refer Slide Time: 09:33)



You will see one more problem according to the National Association of Insurance Commissioners the average annual cost of automobile insurance in the United States in a recent year was 691 dollar. Suppose the automobile insurance cost are uniformly distributed in the United States with an average of, from \$200 to 1182 dollar what is the standard deviation of this uniform distribution. So, we have to find out standard deviation of this distribution. Before that will check the mean the mean is given 691 dollar.

So, we will verify this uniform dot mean loc starting point is 200 the difference is the scale is 982 that is 1182 minus sorry \$200, this is 1 so it is the extra 691 dollar if you want to know the standard deviation of the uniform distribution because this formula is different it is not simple standard deviation so uniform dot std loc is a 200, scale is 982 you will get the standard deviation of 283.47.

Next we will move to the normal distribution here also I have inserted a picture of normal distribution you see that, the picture is taken from this link actual exclamation square break here that link okay when you execute this you will get here picture of probability distribution, that picture. First we will have to import a library norm that is imported from scipy so from scipy dot

stats import norm that is the value, mean, standard deviation: 68, 65.5, 2.5 suppose, if x equal to 68 the mean of that normal distribution is 65.5 standard deviation is 2.5 what is the probability? So, we will run that, first you have to run this also, yes the probability is 0.8413.

If you want to x less than that value suppose, if you want to know cumulative distribution of x greater than value you have to subtract from 1. Suppose if we want to move the value 68 and above. So, already known we know up to 68 this much value so, the remaining area is because we know the area of the normal distribution is 1. So, 1 minus remaining that value will give you the right side area. Suppose if you want to know the value between x_1 and x_2 .

For example value 1 less than or equal to x less than or equal to value 2 so it is a very simple printout norm dot cdf you will find out the upper limit and the lower limit. Simply you type the lower limit because the value is already I have declared. Now suppose the between 68 and 63, x values 60 and 63 if we want to know the area that it plays a very simple reason to receive a lot of our time. Suppose what is the probability of obtaining a score greater than 700 on your GMAT test that has mean 494 and standard deviation of 100 assume GMAT score are normally distributed.

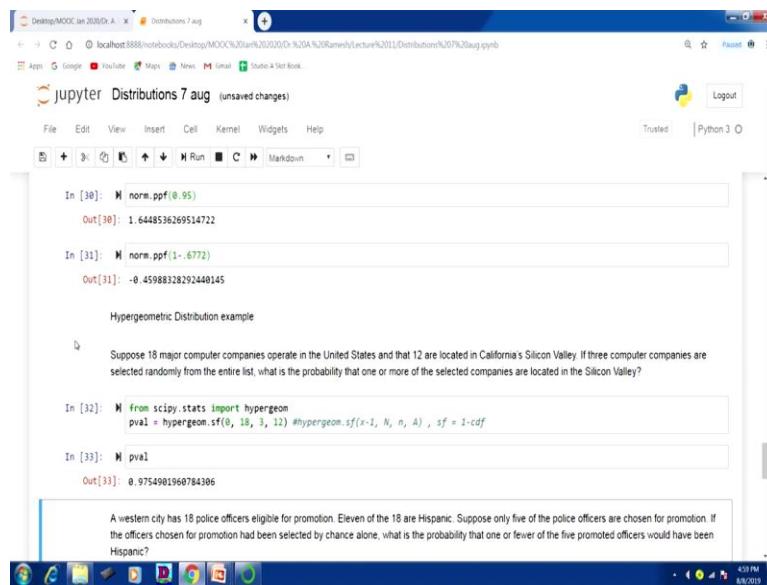
There is another example what is the probability of x greater than 700 when mean equal to 494 and standard deviation is 100. So, because we want to know x greater than equal to 700 so we have to find out x equal to 700 then subtract from 1 so, print 1 minus norm dot cdf 700 the 494, 100 will give you the answer for. what is the probability that randomly drawing his score the 550 or less so we have to need x value less than or equal to 550 so 515, 494, 100 will be the answer.

What is the probability of randomly obtaining a score between 300 to 600 the GMAT examination actually this problem is taken from statistics for management Levin and Reuben. Now you see that the upper limit is 600 lower limit is 300 between 600 and 300 what is the probability so print norm.cdf(600, 494, 100) minus this was upper limit, minus norm.cdf(300, 494, 100) is the lower limit.

What is the probability of getting a score between 350 and 450 on the same GMAT exam, 450 350 there is another example, similar to previous one into this one. Now we are going to do the reverse of that. Now if they are so far be able to find the cdf cumulative probability. Now suppose the area is given if the area is given we want to know the x value, if it is a standard normal distribution we want to know the z value because the default function is the standard normal distribution where the mean equal to 0 standard deviation 1.

So area under 0.95 the corresponding z value is 1.645 this value you can read it from the table the same way which I have explained in the in my theory lecture. Suppose if you want to know most importantly here norm dot ppf that is a probability function. So, now I am norm ppf 1 - 0.672 will give you the left side area. So, we will see what is the corresponding let us say it is z value is, yeah here we are going in the left hand side so the z value is minus 0.459.

(Refer Slide Time: 15:31)



The screenshot shows a Jupyter Notebook window titled "jupyter Distributions 7 aug (unsaved changes)". The notebook has two cells displayed:

```
In [30]: norm.ppf(0.95)
Out[30]: 1.6448536269514722
```

```
In [31]: norm.ppf(1-0.672)
Out[31]: -0.45988328292440145
```

Below the notebook, a text box contains a hypergeometric distribution example:

Suppose 18 major computer companies operate in the United States and that 12 are located in California's Silicon Valley. If three computer companies are selected randomly from the entire list, what is the probability that one or more of the selected companies are located in the Silicon Valley?

Cell In [32] contains the following code:

```
In [32]: from scipy.stats import hypergeom
pval = hypergeom.sf(0, 18, 3, 12) #hypergeom.sf(x-1, N, n, A), sf = 1-cdf
```

Cell In [33] shows the result:

```
In [33]: pval
Out[33]: 0.9754901960784306
```

A text box below the code cells contains a hypergeometric distribution example:

A western city has 18 police officers eligible for promotion. Eleven of the 18 are Hispanic. Suppose only five of the police officers are chosen for promotion. If the officers chosen for promotion had been selected by chance alone, what is the probability that one or fewer of the five promoted officers would have been Hispanic?

Now we will see an example of hyper geometric distribution. the example says suppose 18 major computer companies operate in the United States and that 12 are located in California's Silicon Valley. If 3 computer companies are selected randomly from their entire list what is the probability that one or more of the selected companies are located in the Silicon Valley. What things you have to notice here is 1 or more. So, for that means we have to see what is the probability of getting 1 or more selected companies.

So from scipy dot stats import hypergeom p value equal to hypergeom dot sf, sf means survival function. So, here if it is one or more means that 1 - 1 so 0. 0, 18 represents the population size 3 means we are 3 we are choosing that is the number of sample 3, 12 means the number of success in the population that is a capital A, the same notation what you are used in our theory. So, here the p value is 0.9754.

(Refer Slide Time: 16:45)

```

jupyter Distributions 7 aug (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
In [38]: In [38]: %pylab inline
          from scipy.stats import hypergeom
          pval = hypergeom.sf(0, 18, 3, 12) #hypergeom.sf(x-1, N, n, A), sf = 1-cdf
Out[39]: Out[39]: 0.9754901960784306

A western city has 18 police officers eligible for promotion. Eleven of the 18 are Hispanic. Suppose only five of the police officers are chosen for promotion. If the officers chosen for promotion had been selected by chance alone, what is the probability that one or fewer of the five promoted officers would have been Hispanic?

In [35]: In [35]: pval = hypergeom.cdf(1, 18, 5, 11)
          [1]
Out[36]: Out[36]: 0.04738562091593275

In [ ]: In [ ]: cdf = 1 - pval

```

We will see another example a western city has 18 police officers eligible for promotion 11 of 18 are Hispanic, suppose only 5 of the police officers are chosen for promotion if the officer chosen for promotion had been selected by chance alone what is the probability that one or fewer of the 5 promoted officers would have been his Hispanic. So, what we need to know here 1 or fewer, so here we have to find out the cumulative probability. So, the formula for finding the cumulative probability for a hyper geometric function is the p value.

I am going to save in the name of p value equal to hypergeom dot cdf 1, so 18 represents the population size 5 represents, because choosing 5, 11 represents the number of success in the population. So, when you run this getting 0.04738.

(Refer Slide Time: 17:43)

```

Desktop/MOOC Jan 2020/Dr A -> Distributions 7 aug <+>
localhost:8888/notebooks/Developer/MOOC%20Jan%202020/Dr.%20A.%20Ramey/lectures%2011/Distributions%20%20aug.ipynb
File Edit View Insert Cell Kernel Widgets Help
Logout Trusted Python 3.0
In [40]: In [40]: pval = hypergeom.cdf(1, 18, 5, 11)
In [36]: pval
Out[36]: 0.04738562091583275
Exponential Distribution Example
A manufacturing firm has been involved in statistical quality control for several years. As part of the production process, parts are randomly selected and tested. From the records of these tests, it has been established that a defective part occurs in a pattern that is Poisson distributed on the average of 1.38 defects every 20 minutes during production runs. Use this information to determine the probability that less than 15 minutes will elapse between any two defects?
In [ ]: mu1 = 1/1.38 # for 20 mins
mu1
In [ ]: from scipy.stats import expon
expon.cdf(0.75,0,(1/1.38)) # 15/20 = 0.75, loc=0 because y = (x - loc) / scale, and y= x/scale,
We can also define the function manually
In [ ]: def CDFExponential(lamb,x): #lamb = Lambda

```

Now we will go for next example on exponential distribution we will take a sample problem. A manufacturing firm has involved in statistical quality control for several years. As part of the production process parts are randomly selected and tested from the records of these tests it has been established that the defective part occur in a pattern that is a Poisson distributed on the average of 1.38 defects every 20 minutes during production run. Use the information to determine the probability of less than 15 minutes will elapse between any 2 defects.

Here how to look at the 2 things in this problem one is the mean of your Poisson distributions given mu and second thing is the between any 2 defects. Now when as I told you in theory itself whenever the between 2 things you have to go for exponential distribution. Now first we do find the mean of your exponential distribution. So, the mean of your exponential distribution is 1 by mean of the Poisson distribution. So, here is Poisson distribution mean is 1.38 so the lambda we can call it as mu₁ that is mu₁ is for the mean of here exponential distribution mu₁ equal to 1 divided by 1.38.

So that value is this much suppose, what was asked probability that is less than 15 minutes from scipy we have to import exponential function. So, we have to find out the cumulative probability further to expon dot cdf so the 0.75 represents because we got 0.75 from 15 divided by 20 because that is the mean was in the Poisson distribution mean was for 20 minutes. Now the problem for the exponential distribution is asked for 15 minutes.

So, we are dividing 15 by 20 so that the units are matching. So, we need to find out the cumulative function of exponential distribution. The lower limit of that x is 0 the upper limit is 0.75 so exponent dot cdf upper limit 0.75, lower limit and the lambda value so you will get the 0.644. Students, is so far we have seen we have seen binomial distribution, how to use Python. Then you have seen Poisson distribution, we have seen uniform distribution.

We have seen normal distribution and exponential distribution and hypergeometric distribution also. So, we will continue in the next class with a new topic that is on sampling and sampling distribution, thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 12
Sampling and Sampling Distribution

Dear students we are going to the next lecture that is sampling and sampling distributions. The objective here is; objective of this lecture is describing a simple random sample and why sampling is important.

(Refer Slide Time: 00:40)

Lecture Objectives

After completing this lecture, you should be able to:

- Describe a simple random sample and why sampling is important
- Explain the difference between descriptive and inferential statistics
- Define the concept of a sampling distribution
- Determine the mean and standard deviation for the sampling distribution of the sample mean,

2

Explain the difference between descriptive and inferential statistics and defining the concept of sampling distribution. Determining the mean and standard deviation of the sampling distribution of the sample mean that very important theorem that we are going to see in this class, the central limit theorem and its importance. and determining the mean and standard deviation of the sampling distribution of the sample proportions, then at the end we will see the sampling distribution of sample variances.

(Refer Slide Time: 01:17)

Descriptive vs Inferential Statistics

- **Descriptive statistics**
 - Collecting, presenting, and describing data
- **Inferential statistics**
 - Drawing conclusions and/or making decisions concerning a population based only on sample data



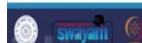
4

The whole statistics can be classified into 2 categories one is the descriptive statistics another one is the inferential statistics. The descriptive statistics is only for collecting and presenting describing the data as it is it is very low-level statistics. Whereas the inferential statistics drawing conclusions are making decisions concerning a population based on sample data, in the inferential statistics with the help of sample data we are going to infer something about the population. So, when you say population you should know what is the population what is the sample?

(Refer Slide Time: 01:58)

Populations and Samples

- A **Population** is the set of all items or individuals of interest
 - Examples: All likely voters in the next election
All parts produced today
All sales receipts for November
- A **Sample** is a subset of the population
 - Examples: 1000 voters selected at random for interview
A few parts selected for destructive testing
Random receipts selected for audit

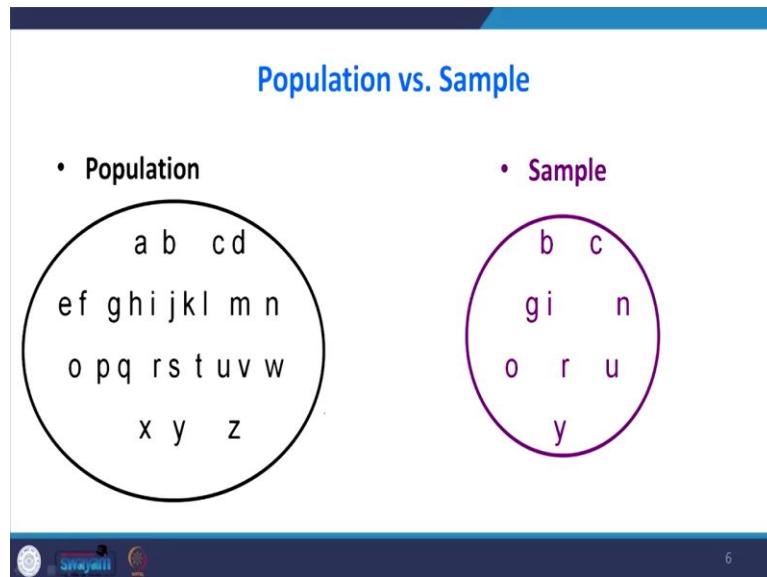


5

Population is the set of all items are individual of interest for example all likely voters in the next election, all parts produced today, all sales received for November. The sample is the subset of

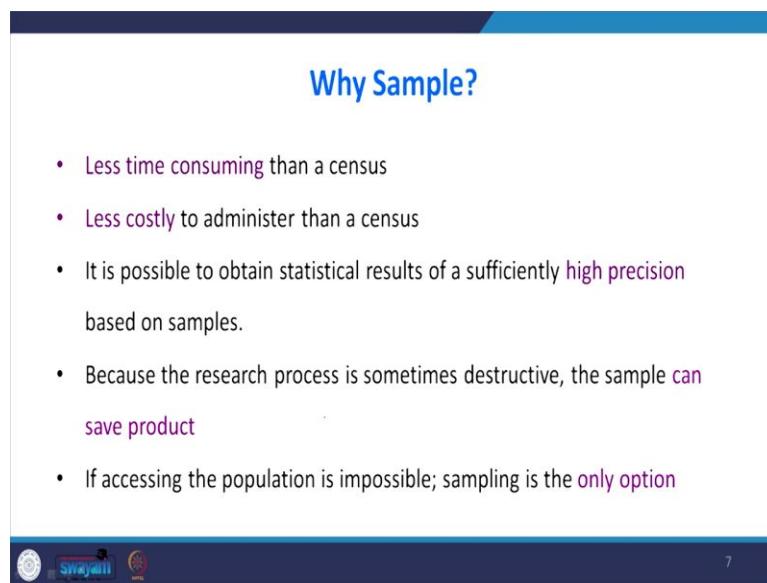
the population like 1000 voters selected at random for interview, a few parts selected for destructive testing, random received selected for audit. This is an example of sample.

(Refer Slide Time: 02:33)



When you look at the left hand side there is a bigger circle that is the population from there some numbers are bigger the collection of that picked of the values is called a sample.

(Refer Slide Time: 02:35)



The question may come why we out to sample it is less time-consuming than a census, less costly to administer then your census. It is possible to obtain statistical result of your sufficient the high precision based on the samples. Because of the research process sometimes destructive the sample can save the product. If accessing the population is impossible sampling is the only

option. Sometimes you have to go for census also we are in census we will examine each and every item in the population.

(Refer Slide Time: 03:17)

Reasons for Taking a Census

- Eliminate the possibility that a random sample is not representative of the population
- The person authorizing the study is uncomfortable with sample information



Suppose if we need to have higher accuracy and you are not comfortable with the sample data then used to go for census. The reasons for taking a census because census eliminates the possibility that random sample is not representative of the population, many time there is a chance that the sample which you have taken may not represent the population. Otherwise the person authorizing the study is uncomfortable with the sample information then you go for census.

(Refer Slide Time: 03:40)

Random Versus Nonrandom Sampling

- **Random sampling**
 - Every unit of the population has the same probability of being included in the sample.
 - A chance mechanism is used in the selection process.
 - Eliminates bias in the selection process
 - Also known as probability sampling
- **Nonrandom Sampling**
 - Every unit of the population does not have the same probability of being included in the sample.
 - Open the selection bias
 - Not appropriate data collection methods for most statistical methods
 - Also known as non-probability sampling



We will see what is sampling? Sampling is generally selecting some items from the population that is a sampling. So, there are that can be classified into two way one is random sampling another one is a non random sampling in the random sampling. The concept of randomness is taken care non random sampling the randomness is not there. Sometimes we may go for non random sampling even though it is not so comfortable that is not good for doing many statistical analysis, sometimes we have to go for non random sampling.

But in the random sampling the outcome or the generalization which you provide with the help of random samplings are highly robust. So, we will go for what is the random sampling? Every unit of the population has the same probability of being included in the sample that is the concept of your randomness. A chance mechanism is used to selection of the process because the chance of mechanism is we can use a random table to choose someone, you can use your calculator you can choose someone, choose someone randomly that eliminates the bias in the selection process also known as the probability sampling.

They will go for non random sampling every unit of the population does not have the same probability of being included in the sample. It is open the, you know selection bias, there is a possibility selection bias not appropriate data collection methods for most statistical methods. So, it is not good method for doing some statistical analysis also known as non-probability sampling.

(Refer Slide Time: 05:18)

Random Sampling Techniques

- Simple Random Sample
- Stratified Random Sample
 - Proportionate
 - Disproportionate
- Systematic Random Sample
- Cluster (or Area) Sampling

Random sampling techniques there are 4 way we can say of selecting random one is the simple random sample second one is a stratified random sample with the proportion disproportionate third one is a systematic random sample fourth one is cluster or area sampling. Simple random samples every object in the population has an equal chance of being selected, objects are selected independently.

(Refer Slide Time: 05:48)

Simple Random Samples

- Every object in the population has an **equal chance** of being selected
- Objects are selected independently
- Samples can be obtained from a table of random numbers or computer random number generators
- A simple random sample is the ideal against which other sample methods are compared



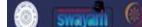
11

Samples can be obtained from your table of random numbers or computer random number generators. A simple random sample is the ideal against which the sample methods are compared this is a best method.

(Refer Slide Time: 05:59)

Simple Random Sample: Numbered Population Frame

01 Andhra Pradesh 02 Himachal Pradesh 03 Gujrat 04 Maharashtra 05 Nagaland 06 Goa 07 West bengal 08 Haryana 09 Punjab 10 Delhi	11 Madhya Pradesh 12 Uttar Pradesh 13 Bihar 14 Rajasthan 15 J & K 16 Tamil Nadu 17 Karantaka 18 Kerala 19 Orissa 20 Manipur
---	--



Suppose we will see there are 20 states have ever taken suppose I want to choose some states randomly for some studies. Suppose first task is I have given some number 2-digit number 01, 02 for example up to this one, it is only for illustrate the purpose it is not 20 the number of states are more.

(Refer Slide Time: 06:22)

The slide has a dark blue header bar with the title "Simple Random Sampling: Random Number Table". Below the title is a 10x10 grid of random numbers. The first two rows of the grid are highlighted with a yellow border. The numbers in the grid are as follows:

9	9	4	3	7	8	7	9	6	1
4	5	7	3	7	3	7	5	5	2
9	7	9	6	9	3	9	0	9	4
3	4	4	7	5	3	4	4	7	5
3	1	6	1	8					
5	0	6	5	6	0	1	2	7	6
6	8	3	6	7	6	6	8	2	0
8	0	8	8	0	8	1	5	6	8
6	3	1	7	1	4	2	8	7	7
8	6	8	3	5	6	0	5	1	5
8	6	4	2	0	4	0	5	1	5
8	6	4	8	5	3	5	3	5	9
8	6	4	2	0	4	0	0	9	4
8	6	4	3	6	0	1	8	6	9
8	6	4	3	6	0	1	8	6	9
8	6	4	3	6	0	1	8	6	9
5	2	5	8	7	1	9	6	5	3
7	1	9	6	5	8	5	4	5	3
8	5	4	5	3	4	6	8	3	4
8	9	1	5	5	3	0	0	9	9
9	0	5	5	3	9	0	6	8	9
9	0	6	8	9	4	8	6	3	7
0	7	9	5	5	4	7	0	6	2
0	7	9	5	5	4	7	0	6	2
7	6	9	4	8	1	0	4	9	3
7	6	9	4	8	1	0	4	9	3
7	6	9	4	8	1	0	4	9	3
7	6	9	4	8	1	0	4	9	3

Next I am using the random table to choose the States randomly. For example you can start from you can see this is a random table you say see the table you can follow any 2 digit 99, 43, 78, 79, 61, because the random table can be read at any direction. so suppose if I am reading left to right 99, 43, 78, 76, 61, 45 and so on so 53 next is 16 so 16 is, I have to choose the serial number 16 and corresponding states I am going back so the 16 is Tamilnadu. So, one state is chosen the next random number is 18 so the 18 is Kerala next state is chosen.

Next 50 there is no number 65 there is no number 60 there is no number but 01 number is a 01 is Andhra Pradesh then 27, 27 is not there 68 not there's 36 not there 76 not there 68 not there 82 is not but 08 is there 08 it is Haryana. So, like this, this is the way to use a random table to choose from the population. Here the population is the number of states suppose I want to choose some states randomly for my study so I can use the, this random number table.

Suppose so the capital N is a 20 n is 4 so capital N represents the population n represents the sample size.

(Refer Slide Time: 08:08)

Stratified Random Sample

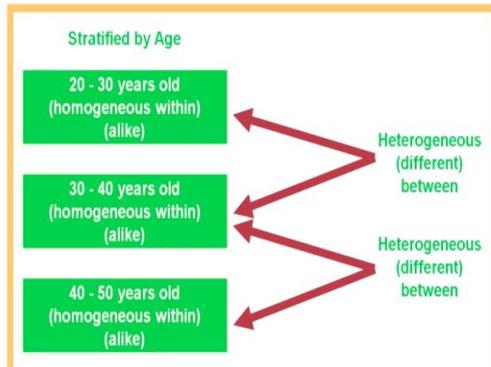
- Population is divided into non-overlapping subpopulations called strata
- A random sample is selected from each stratum
- Potential for reducing sampling error
- Proportionate -- the percentage of these samples taken from each stratum is proportionate to the percentage that each stratum is within the population
- Disproportionate -- proportions of the strata within the sample are different than the proportions of the strata within the population



Then we will go for stratified sampling so the population is divided into non-overlapping subpopulations called strata. Random sample is selected from each stratum potential for reducing sampling error. We can go for proportionate the percentage of these samples taken from each stratum is proportion to the percentage that each stratum is within the population. We can go for disproportionate also the proportion of strata within the sample are different than the proportion of the strata within the population.

(Refer Slide Time: 08:41)

Stratified Random Sample: Population of FM Radio Listeners



For example stratified random sample population of FM radio listeners so what I have done the whole population is divided into 3 stratum one is 20 to 30, 30 to 40, 40 to 50 you see that each

stratum are homogenous within between the stratum there may be a difference maybe there is a different variance but the same stratum will have is homogeneous the similar kind of behavior are dataset it will be there. Why it is reducing the sampling error that you if you choose 20 to 30 if you choose something from this strata so all we will we have the similar characteristics.

If you choose number some numbers 40 to 50 these sample will have similar characteristics. See that between the stratum it is a heterogeneous within the strata it is homogeneous.

(Refer Slide Time: 09:39)

Systematic Sampling

- Convenient and relatively easy to administer
- Population elements are an ordered sequence (at least, conceptually).
- The first sample element is selected randomly from the first k population elements.
- Thereafter, sample elements are selected at a constant interval, k , from the ordered sequence frame.

$$k = \frac{N}{n},$$
 where:
 n = sample size
 N = population size
 k = size of selection interval



Then next method is the systematic sampling it is convenient and relatively easy to administer the population elements are ordered in sequence. The first sample element is selected randomly from the first K population element. Thereafter the sample elements are selected at a constant interval k from the ordered sequence of frame. What is the k is, k is the population size divided by sample size. The k represents the size of selection interval we will see an example.

(Refer Slide Time: 10:12)

Systematic Sampling: Example

- Purchase orders for the previous fiscal year are serialized 1 to 10,000 ($N = 10,000$).
- A sample of fifty ($n = 50$) purchases orders is needed for an audit.
- $k = 10,000/50 = 200$
- First sample element randomly selected from the first 200 purchase orders. Assume the 45th purchase order was selected.
- **Subsequent sample elements: 245, 445, 645, ...**



Suppose the purchase order is from the previous fiscal year serialized one to 10,000 so capital N is 10,000 a sample of 50 n equal to 50 purchases orders need to be selected for an audit so here K is 10,000 divided by 50 that is a 200, K is the interval so the first sample element randomly selected from the first 200 purchases assuming that we have chosen 45th the purchase order from the 45th you have to add 200, so 45th plus 200, 245, 245 +, 445+ 645, and so on.

(Refer Slide Time: 10:12)

Cluster Sampling

- Population is divided into non-overlapping clusters or areas
- Each cluster is a miniature of the population.
- A subset of the clusters is selected randomly for the sample.
- If the number of elements in the subset of clusters is larger than the desired value of n , these clusters may be subdivided to form a new set of clusters and subjected to a random selection process.



Then we will go for the cluster sampling here the population is divided into non-overlapping clusters or areas. Each cluster is miniature of the population the subset of cluster is selected randomly from the sample if the number of elements in the subset of cluster is larger than the desired value of n these clusters may be subdivided into form a new set of clusters and subjected

to a random selection process. Because each cluster will behave like your population now you may ask the difference between stratified sampling and cluster sampling.

In stratified sampling the things are homogeneous in each stratum the items within the in Stratham of homogenous but in cluster sampling it is highly heterogeneous and each cluster will act like your population. For example say, apparel cluster Ludhiana, apparel cluster Tirupur or these are the example of clusters because each cluster will have similar characteristics but will have different variants.

(Refer Slide Time: 12:04)

Cluster Sampling

- ◆ **Advantages**
 - More convenient for geographically dispersed populations
 - Reduced travel costs to contact sample elements
 - Simplified administration of the survey
 - Unavailability of sampling frame prohibits using other random sampling methods
- ◆ **Disadvantages**
 - Statistically less efficient when the cluster elements are similar
 - Costs and problems of statistical analysis are greater than for simple random sampling

So, we will go for advantages of cluster sampling it is more convenient for geographically dispersed a population, reduced travel cost to contact the sample elements, simplify the administration of the survey because the cluster itself will act as a population. Unavailability of sampling frame prohibits using other random sampling methods because there is no other method we can go for a cluster sampling. The disadvantage is statistically less efficient when the cluster elements are similar.

Because that cannot be generalized cost and problem of static analysis are greater than simple random sampling.

(Refer Slide Time: 12:40)

Nonrandom Sampling

- **Convenience Sampling:** Sample elements are selected for the convenience of the researcher
- **Judgment Sampling:** Sample elements are selected by the judgment of the researcher
- **Quota Sampling:** Sample elements are selected until the quota controls are satisfied
- **Snowball Sampling:** Survey subjects are selected based on referral from other survey respondents



The next kind of sampling technique is non-random sampling the first one is the convenience sampling because based on the convenience of the researcher the sample is selected. Next one is the judgement sampling sample elements are selected by the judgement of the researcher for example suppose you administering a questionnaire suppose that questionnaire can be understood only by a manager then you have to look for only the managers. So, the researcher is judging that who should fill this questionnaire so judgment sampling.

Then quota sampling sample elements are selected until quota controls are stratified. Suppose say some, Uttarakhand there are some districts and each distinct I have to collect some sample so I may have some quota for example in Haridwar district how much sample has to be collected some other district how many sample has to be collected. So, there is a quota sampling. Snowball sampling is a very familiar that survey objects are selected based on the referral from other survey respondents.

Suppose you may approach one respondent out ever the survey is over you can ask him to refer his friends, so that is a snowball sampling. It is a very common method in the research.

(Refer Slide Time: 13:56)

Errors

- Data from nonrandom samples are not appropriate for analysis by inferential statistical methods.
- Sampling Error occurs when the sample is not representative of the population
- Non-sampling Errors
 - Missing Data, Recording, Data Entry, and Analysis Errors
 - Poorly conceived concepts , unclear definitions, and defective questionnaires
 - Response errors occur when people do not know, will not say, or overstate in their answers



Then there are some errors when we go, when we go for sampling. Data from non-random samples are not appropriate for analysis of inferential statistical methods that was there a very important drawback because you cannot generalize because there is no randomness. Sampling error occurs when the sample is not the representative of the population, if the sample is not representing the population then whatever analysis you do that will become futile.

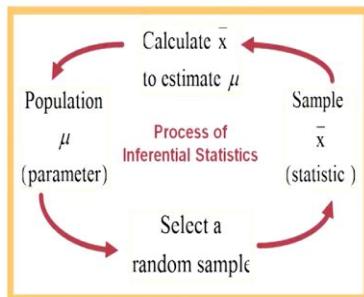
So, non sampling error suppose if you go for apart from this sampling procedure sometime there may be missing data, there may be problem and recording, there may be problem with the data entry, there may be analysis error. Sometime the poorly consumed concepts, unclear definition and defective questionnaires that also lead to error. Sometime response error occurs when the people may not understand what is the questionnaire.

Suppose there is option that not know, will not say. Sometimes the respondent may over state their answers, these are the possible error when you go for sampling. There is one more error, type 1 and type 2 error that we will see in the coming classes.

(Refer Slide Time: 15:19)

Sampling Distribution of \bar{X}

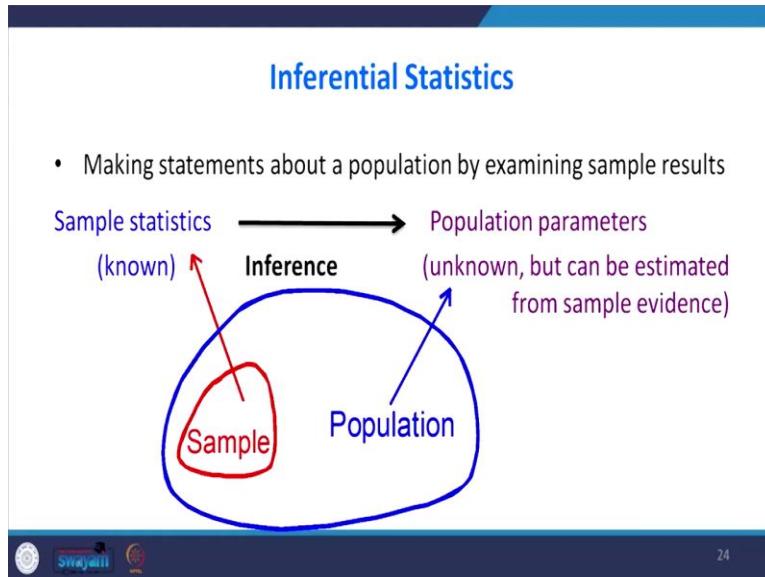
Proper analysis and interpretation of a sample statistic requires knowledge of its distribution.



So, now is to go to the sampling distribution of mean here \bar{X} represents the mean so the proper analysis and interpretation of your sample statistic require knowledge of its distribution that is a sampling distribution. For example we start from population say population is μ select a random sample from the sample you select the sample statistic, statistic it is not statistics, yes there is no s, so whatever things would you say about the sample it is called a statistic, T statistic F statistic \bar{X} -bar these are; since we you calculated from the sample we are calling it to statistic.

With the help of sample mean you can calculate or estimate the operation mean this is the process of we were inferential statistics. So, what is happening something we are going to assume about the population once we assume that population that is generally called hypothesis then we will take a sample randomly we will do some sample statistic with the help of sample statistic we can estimate the population mean or we can estimate the population variance. In this contest currently we are estimating the population mean.

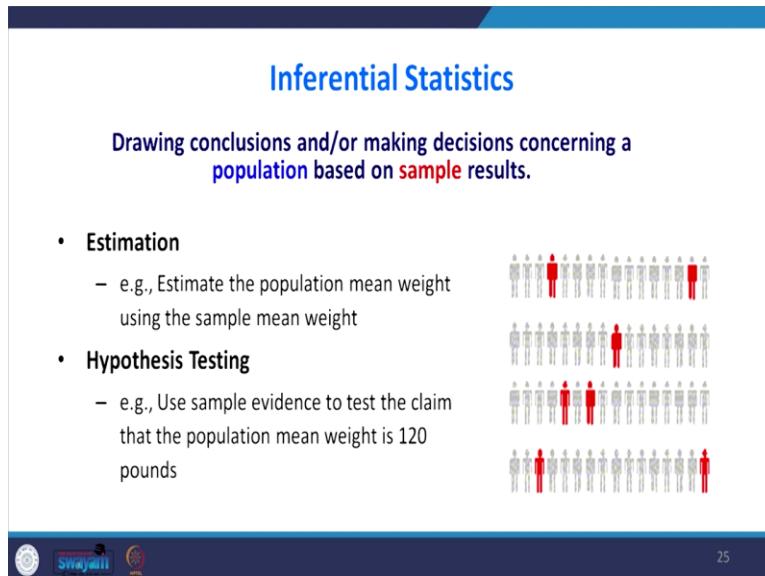
(Refer Slide Time: 16:37)



24

This picture shows the inferential statistics there is,m see there are bigger circle that is the population. So, the population parameter is unknown but can be estimated from the sample evidence see the red one shows that the sample statistic. So, what is the inferential statistics is making statements about a population by examining sample result that is the inferential statistic.

(Refer Slide Time: 17:04)



25

See another example of inferential statistics drawing conclusions or making decision concerning a population based on these sample results. You see there are different red color is there. So, these 1 2 3 4 5 6 7 these are the sample the whole things in the population, the inferential statistics is used for estimation estimating the population mean weight using the sample mean weight. For example if you want to know the weight of the population that can be estimated with

the help of weight of the sample mean then this inferential statistics are another application was for hypothesis testing.

We can use sample evidence to test the claim that the population mean weight is for example 120 pounds are not. We will go in detail about the statistics in coming lectures.

(Refer Slide Time: 17:57)

Sampling Distributions

- A **sampling distribution** is a distribution of all of the possible values of a statistic for a given size sample selected from a population

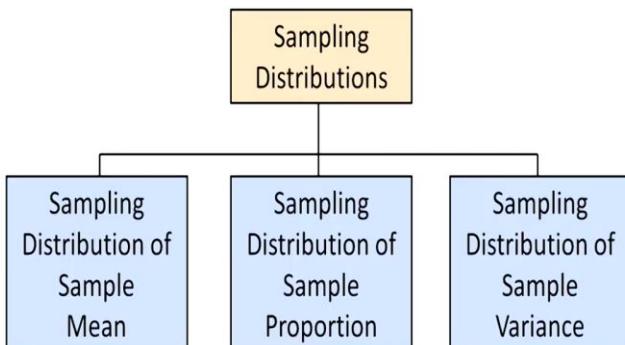


26

Now we are entering into the sampling distribution sampling distribution is a distribution of all of the possible values of your statistic for a given size sample selected from the population.

(Refer Slide Time: 18:14)

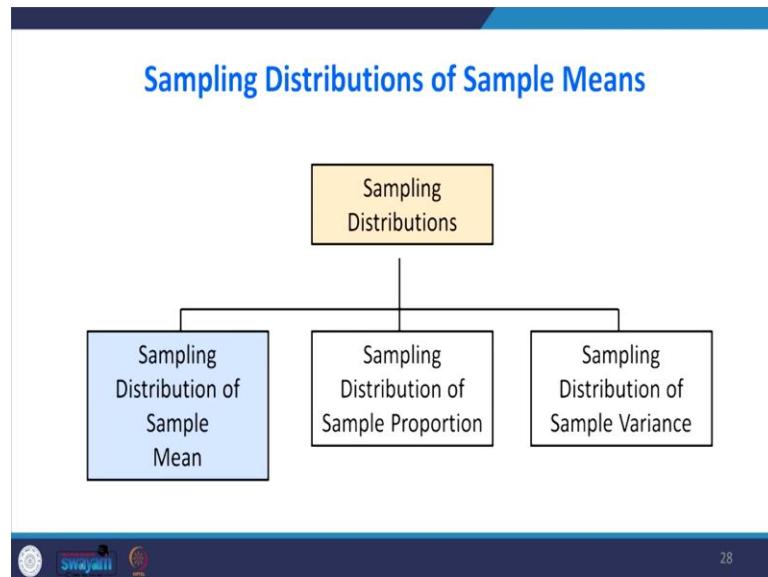
Types of sampling distributions



27

So, what will happen we can say type of sampling distributions we can do the sampling distribution for the sample mean. We can do the sampling distribution for the sample proportion. We can do the sampling distribution for sample variance.

(Refer Slide Time: 18:30)



First we will see the sampling distribution of sample mean.

(Refer Slide Time: 18:34)

Assume there is a population ...
Population size $N=4$
Random variable, X , is age of individuals
Values of X : 18, 20, 22, 24 (years)

A B C D

29

Suppose assume that there is a population there are 4 people in a population that is age random variable is x is age of individuals. So, the value of x may be 18, 20, 22, 24 it is the population.

(Refer Slide Time: 18:54)

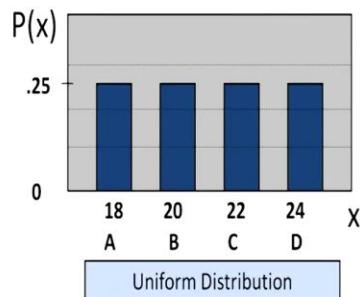
Developing a Sampling Distribution (continued)

Summary Measures for the Population Distribution:

$$\mu = \frac{\sum X_i}{N}$$

$$= \frac{18 + 20 + 22 + 24}{4} = 21$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = 2.236$$



First you will find out the population mean population mean is Sigma of capital X_i divided by N generally whenever you see a capital alphabet that is for the population. The smaller one is for the variance. So, 18, 20, 22, 24 divided by 4 is 21 similarly the population variance is 2.236. What is happening there are 4 element is there so the probability of getting each element that is choosing 18, 20 it is 1 by 4 so 0.25 + 0.25 and 0.25 it this follow uniform distribution.

Suppose if we choose only one sample when you plot it the chances for selecting each person from the population is 0.25.

(Refer Slide Time: 19:49)

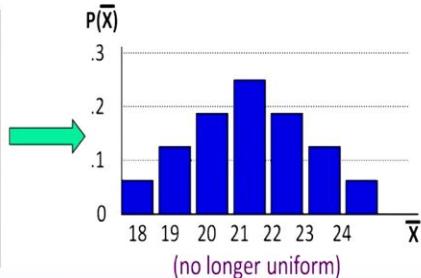
Developing a Sampling Distribution (continued)

- Sampling Distribution of All Sample Means

16 Sample Means

1st Obs	2nd Observation	18	20	22	24
18	18	19	20	21	
20	19	20	21	22	
22	20	21	22	23	
24	21	22	23	24	

Sample Means Distribution



Suppose if you consider all possible sample of size n, size n here means we are going to select 2 people with the replacement there is a possibility first observation may be 18 20 22 24 second observation may be 18 20 22 24 so possibility is 18 18, 18 20, 18 22, 18 24, 20 18, 20 20, 20 22 and so on. So, there are 16 possible samples here we are doing sampling with replacement that is why it is coming 20 20, 22 22, 24 24.

If we find the mean of this, so right side picture shows the mean of that 18 18 is 20, 18 20, 19 when you plot this me what is happening that mean of this sample is following normal distribution. |Previously when you take only one sample when you plot it we are getting uniform distribution. When you increase the sample size 1 to 2 what is happening you are getting here normal distribution it is no longer uniform.

(Refer Slide Time: 21:00)

Developing a Sampling Distribution *(continued)*

- Summary Measures of this Sampling Distribution:

$$E(\bar{X}) = \frac{\sum \bar{X}_i}{N} = \frac{18+19+21+\dots+24}{16} = 21 = \mu$$

$$\sigma_{\bar{X}} = \sqrt{\frac{\sum (\bar{X}_i - \mu)^2}{N}}$$

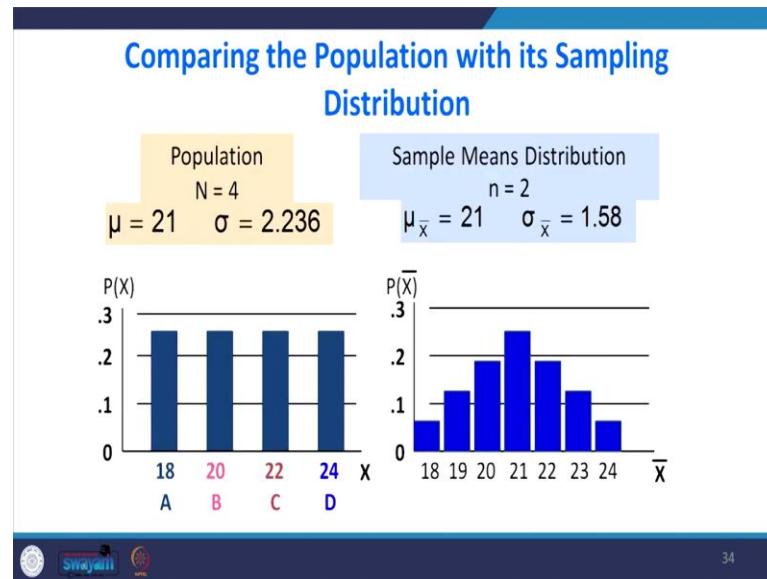
$$= \sqrt{\frac{(18-21)^2 + (19-21)^2 + \dots + (24-21)^2}{16}} = 1.58$$

Now summary measure of this sampling distribution where we selected to with replacement you see that and going back there are 16 elements 4 x 4, 4 times 4 =16 element. So, the mean expected value of x bar is 18 19 21 up to 24 out of 16, mu equal to 21. Then the standard deviation of this sampling distribution is $\Sigma(X - \mu)^2 / \sqrt{N}$, so the formula for standard deviation is first to find the variance, mu is 21, so $(18 - 21)^2 + (19 - 21)^2 + \dots + (24 - 21)^2 =$ it is 1.58.

Please look at and going back look at the population mean. The population mean is 21 and population standard deviation is to 2.236, when we select 2 with replacement mean of the

sampling distribution is 21 but the standard deviation of the sampling distribution is 1.58 when you go for selecting 2 samples with replacement.

(Refer Slide Time: 22:16)

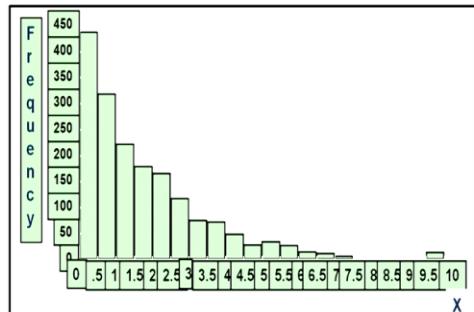


Next what we have to do we are going to select the 4 at a time we are going to construct the same table which have constructed previously. After constructing when you find the mean it will be 21 so we have found these summary measures for the sampling distribution where the mean of the sampling distribution is 21 and the standard deviation of sampling distribution is 1.58, so when we compare population data versus sample.

For population there are 4 element in the population in the sample there are 2 element. The mean of the population is 21 the mean of the sampling distribution is also 21 but the standard deviation of the population is 2.236 but the standard deviation of sample distribution is 1.58.

(Refer Slide Time: 23:08)

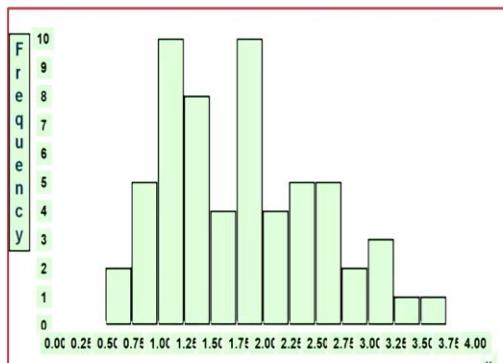
1,800 Randomly Selected Values from an Exponential Distribution



You will go for another example that there is a population which follow an exponential distribution. Now from this exponential distribution we are going to select 2 at a time with replacement. When you select two at a time then if I find the mean then if you construct frequency distribution then if I plot that frequency distribution when n equal to 2 we are getting this kind of distribution you see that the parent distribution is exponential when the sample size is 2 if I plot the mean of the sample mean that is following this kind of similar to uniform distribution.

(Refer Slide Time: 23:50)

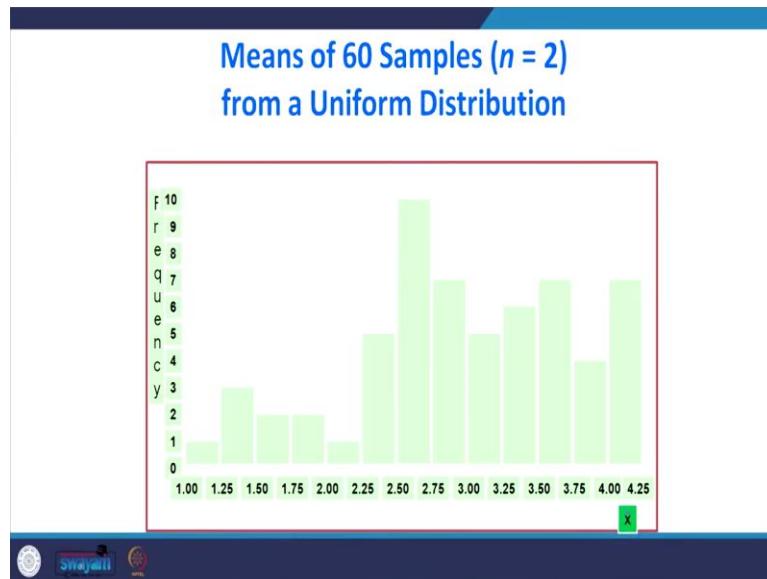
Means of 60 Samples ($n = 5$) from an Exponential Distribution



If I increase the sample size to 5 what is happening it is changed. so when n equal to 30 it is looking like here normal distribution. So, what is happening whatever may be the nature of the

population if you select any sample from the population then if you plot that the sample mean that will follow normal distribution. So, for example another example you take the population follow a uniform distribution.

(Refer Slide Time: 24:21)



You select 2 at a time and plot the sample mean that follow this kind of distribution increase sample size to 5 it is approaching normal distribution. When n equal to 30 it is looking like a normal distribution initially it was the uniform distribution when the sample size is increasing then it is following it is behaving like a normal distribution.

(Refer Slide Time: 24:43)

Expected Value of Sample Mean

- Let X_1, X_2, \dots, X_n represent a random sample from a population
- The sample mean value of these observations is defined as

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

So, expected value of sample mean let $X_1 X_2 \dots X_n$ represent a random sample from the

$$\bar{X} = \frac{\sum X_i}{N}$$

population. The sample mean of these observations is defined as \bar{X} , then standard error of the mean. different samples of the same size from the same population yield different sample means. A measure of variability in the mean from the sample to sample is given by standard error of the mean.

So standard error is σ/\sqrt{n} , note that the standard error of the mean decreases when the sample sizes increases.

(Refer Slide Time: 25:31)

If sample values are not independent (continued)

- If the sample size n is not a small fraction of the population size N , then individual sample members are not distributed independently of one another
- Thus, observations are not selected independently
- A correction is made to account for this:

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \frac{N-n}{N-1} \quad \text{or} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

45

See if the sample values are not independent what will happen if the sample size is n and not a small fraction of the population size capital N then the individual sample members are not distributed independently of one another thus observations are not selected independently. So, a correction is made to account for this. So, σ^2/n that was the variance of the sampling distribution that has to be multiplied by $(N - n / n - 1)$. You take square root of it, $\sigma \cdot \sqrt{(N-n) / (N-1)} \cdot n$.

(Refer Slide Time: 26:09)

If the Population is Normal

- If a population is **normal** with mean μ and standard deviation σ , the sampling distribution of \bar{X} is **also normally distributed** with

$$\mu_{\bar{X}} = \mu \quad \text{and} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- If the sample size n is not large relative to the population size N , then

$$\mu_{\bar{X}} = \mu \quad \text{and} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$



46

Ok if the population is normal with the mean mu and standard deviation Sigma the sampling

$\mu_{\bar{X}} = \mu$
distribution of X is also normally distributed with the

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

,

$$\mu_{\bar{X}} = \mu$$

When the sample size is not a large relative to the population then

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

(Refer Slide Time: 26:36)

Z-value for Sampling Distribution of the Mean

- Z-value for the sampling distribution of :

$$Z = \frac{(\bar{X} - \mu)}{\sigma_{\bar{X}}}$$

where:
 \bar{X} = sample mean
 μ = population mean
 $\sigma_{\bar{X}}$ = standard error of the mean

47

$$z = \frac{(\bar{X} - \mu)}{\sigma_{\bar{X}}}$$

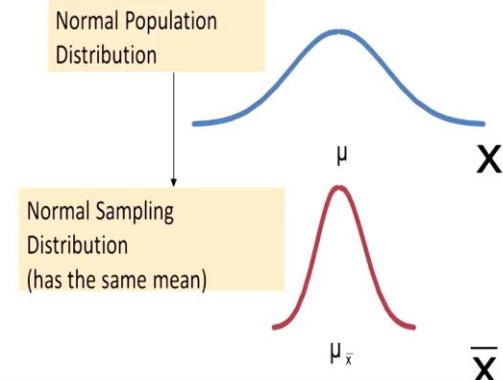
So, the Z value for the sampling distribution of the mean is

(Refer Slide Time: 21:00)

Sampling Distribution Properties

$$\mu_{\bar{X}} = \mu$$

(i.e. \bar{X} is unbiased)



48

We look at the sampling distribution properties, see the top one it is a normal population distribution but the normal sampling distribution has the same mean, then sampling distribution properties. For sampling with replacement when n increases sample size increases the standard deviation of sampling distribution decreases. So, what is happening look at the red color there is a large sample size that is the smaller standard deviation. Look at the blue one smaller sample size larger standard deviation.

(Refer Slide Time: 27:20)

If the Population is not Normal- Central Limit Theorem

We can apply the Central Limit Theorem:

- Even if the population is **not normal**,
- sample means from the population **will be approximately normal** as long as the sample size is large enough.

Properties of the sampling distribution:

$$\mu_{\bar{x}} = \mu \quad \text{And} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

50

The population is not normal we can apply the central limit theorem even if the population is not normal. Sample means from the population will be approximately normal as long as the sample

size is large enough. The properties of sampling distribution is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

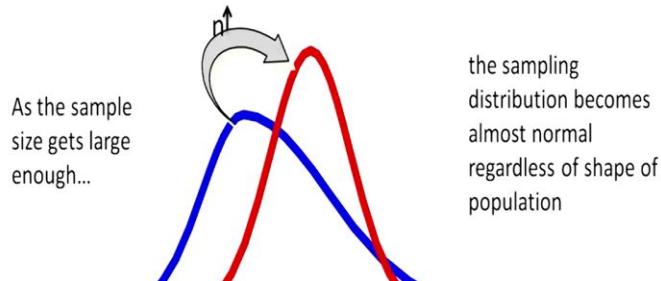
,

. This theorem is very important theorem that is the central limit theorem, why it is important through this theorem the concept of sample and population is connected. What is the result the mean of the sampling distribution is population mean.

The standard deviation of sampling distribution is σ/\sqrt{n} , where the Sigma represents the population standard deviation n represents the sample size. It is very powerful it is the very fundamental theorem for inferential statistics.

(Refer Slide Time: 28:19)

Central Limit Theorem



What is happening as the sample size get large enough the sampling distribution becomes almost normal regardless of the shape of the regardless of shape of the population. So, what is the meaning is suppose there is a population you take some sample if you plot the sample mean that will follow normal distribution provided n is large enough. So, when you keep on increase n then the sampling distribution will be exactly like your normal distribution.

(Refer Slide Time: 28:52)

If the Population is not Normal

(continued)

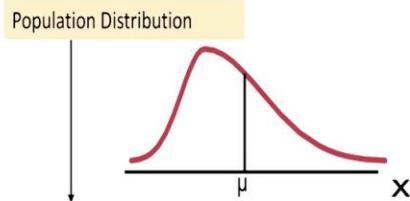
Sampling distribution properties:

Central Tendency

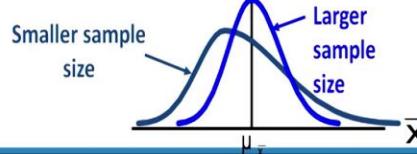
$$\mu_{\bar{x}} = \mu$$

Variation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



Sampling Distribution (becomes normal as n increases)



Even this is applicable even the population is not normal the parent population may be may follow any distribution but the sampling distribution will always will follow a normal distribution so the mu X bar equal to mu the standard deviation is Sigma by root n. You look at

this case also the population is not following a normal distribution but the sampling distribution will follow normal distribution.

(Refer Slide Time: 29:21)

How Large is Large Enough?

- For most distributions, $n > 25$ will give a sampling distribution that is nearly normal
- For normal population distributions, the sampling distribution of the mean is always normally distributed

53

So how large is large enough for most distribution when n is greater than 25 will assembling distribution that is nearly normal. For normal population distributions the sampling distribution of the mean is always normally distributed very important result. What is the meaning the sampling distribution of the mean is always normally distributed.

(Refer Slide Time: 29:46)

Example

- Suppose a large population has mean $\mu = 8$ and standard deviation $\sigma = 3$.
Suppose a random sample of size $n = 36$ is selected.
- What is the probability that the sample mean is between 7.8 and 8.2?

54

Suppose we will see an example a large population has mean equal to 8 and standard deviation is 3 suppose a random sample of $n = 36$ is selected what is the probability that the sample mean is between 7.8 and 8.2, we will see an example.

(Refer Slide Time: 30:06)

Example

Solution:

- Even if the population is not normally distributed, the central limit theorem can be used ($n > 25$)
- ... so the sampling distribution of \bar{X} is approximately normal
- ... with mean $\mu_{\bar{x}} = 8$
- ...and standard deviation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$$



55

Even if the population is not normally distributed the central limit theorem can be used when n is greater than 25. So, the sampling distribution of x -bar is approximately normal that is the result

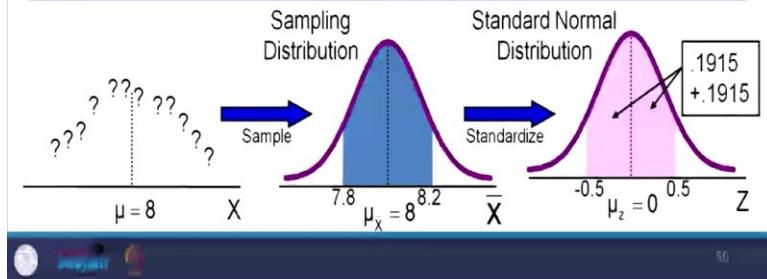
which have seen $\mu_{\bar{x}}$ equal to 8 and this standard because the $\mu_{\bar{x}}$ is mean of the sampling distribution is 8 the standard deviation of the sampling distribution is Sigma by root n, Sigma is 3, n is 36, so 0.5.

(Refer Slide Time: 30:33)

Example (continued)

Solution (continued)

$$\begin{aligned} P(7.8 < \mu_{\bar{x}} < 8.2) &= P\left(\frac{7.8 - \mu}{\sigma/\sqrt{n}} < \frac{\mu_{\bar{x}} - \mu}{\sigma/\sqrt{n}} < \frac{8.2 - \mu}{\sigma/\sqrt{n}}\right) \\ &= P(-0.5 < Z < 0.5) = 0.3830 \end{aligned}$$



So what will happen we were asked $P(7.8 < \mu_{\bar{x}} < 8.2)$ so this 7.8 has to be converted to Z scale the conversion factor the conversion formula from converting to Z it is $X - \mu$ by Sigma by root n the X is given 7.8 mu is 8.2 Sigma is 3 sample sizes 36 that will be the middle one that is $\mu_{\bar{x}} - \mu / (\sigma/\sqrt{n})$ that is nothing but your Z value less than equal to the upper limit.

So X is $8.2 - \mu$ 2 by 3 by root of the 36 so, when you simplify this P of - 0.5 less than Z less than 0.5 that will give you the probability of 0.3830 so what is happening the extreme left shows the picture of your population there is a question mark that means the population may follow any distribution. if you select some sample when you find the sample mean then you draw the sampling distribution that will follow normal distribution.

So what is the area of the sampling distribution between 7.8 and 8.2 that was asked otherwise what is the probability of that the mean of the sampling distribution is between 7.82 to 8.2 so that 7.8 has to be converted into Z scale so that we can refer the table that conversion is done with the help of formulas Z equal to $(X - \mu)$ by σ by root n after converting the 7.8 corresponding Z values - 0.5, 8.2 corresponding Z values 0.5.

We can look at the statistical table or we can use Python to find the area between - 0.5 to 0.5 that will give the area of .1915, + 0.1915 with that we will close this one. So, I'm concluding this lecture so what we have seen in this lecture is different sampling techniques. We have seen the

importance of the sampling then we have seen the probability sampling non-probability sampling. Next we have seen an explanation for our central limit theorem.

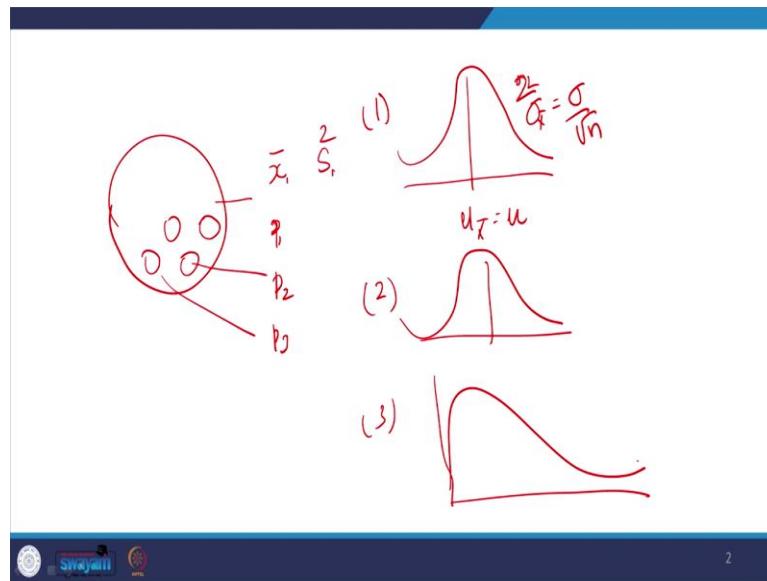
What is the central limit theorem the nature of the population may be anything if you take some sample from the population if you plot that sample mean that will always follow a normal distribution that is the central limit theorem because the central limit theorem is very important theorem we have seen one problem also by using central limit theorem. We will continue in the next class.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 13
Distribution of Sample Mean, Proportion, and Variance

Welcome Students, last class we have started various sampling distributions. We have seen the sampling distribution of mean. In this class, we will continue from the sampling distribution of mean then we will talk about the sampling distribution of proportions and sampling distribution of the variance. Then I will introduce the concept of chi-squared distribution. We will do some problems then with that we can close this lecture. Before going to this lecture just to recollect what we have done in the previous class.

(Refer Slide Time: 00:31)



This is a population from the population I am taking a sample 1, say, sample 2, sample 3, sample 4. For each sample I can find out the sample mean and sample variance for example if I write \bar{x}_1 bar this is for sample 1 its corresponding mean, if I say, s_1^2 it is the sample with sample variance. So what will you do this is for continuous variable. Continuous variable in the sense if I am measuring some length or height or something, suppose if I take the same sample assume that I am taking a discrete variable or the categorical variable, categorical, categorical variable in the sense, it can have only two values positive negative or good or bad.

Suppose I am taking so this is sample one out of the sample 1, e how many good product is there. So, what is the proportion so then I can call it this is P 1, P 1, another sample that will be P 2 another sample that is P 3. So, if I plot this P1, P2, P3 directly I cannot plot it. First I have to construct a frequency distribution then, I have to plot it I will get another distribution that is the sampling distribution of proportion.

So, there are three point here one is first you take the sample, you take the sample mean, if I plot that sample mean that will follow a normal distribution. So, mean of the sampling distribution is

$$\mu_{\bar{x}} = \mu$$

the variance of sampling distribution is I am writing $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ This is my first result. What is the first result? From the population I have taken the sample, if I plot that sample, we will mean that will follow normal distribution.

Similarly, in this lecture, what you are going to do, we are going to take some sample from the population, each population is you know, each sample is going, we are going to find out the proportion so proportion means the probability. So I make it P1, P2, P3 that also will follow normal distribution okay. The third one, which we are going to see in the class, so, we have taken the mean, if you take the variance of each sample, if I plot that variance, if I plot that variance which has come from normal distribution that will follow a special shape, this is called chi-square distribution okay. This is going to be summary of our class. We will continue yeah.

(Refer Slide Time: 04:01)

Acceptance Intervals

Goal: determine a range within which sample means are likely to occur, given a population mean and variance

- By the Central Limit Theorem, we know that the distribution of X is approximately normal if n is large enough, with mean μ and standard deviation
- Let $z_{\alpha/2}$ be the z-value that leaves area $\alpha/2$ in the upper tail of the normal distribution (i.e., the interval $-\infty$ to $z_{\alpha/2}$ encloses probability $1 - \alpha$)
- Then

$$\bar{X} = \mu \pm z_{\alpha/2} \sigma_{\bar{X}}$$

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

is the interval that includes X with probability $1 - \alpha$



3

Before that from the previous class we have seen sampling distribution of the mean, with the help of sampling distribution of the mean, we can find out the lower, upper limit of a sample

$$\mu \pm z_{\alpha/2} \sigma_{\bar{X}}$$

mean that is done with the help of .

We will see how it is Goal: Determine a range within which the sample means are likely to occur given a population mean and variance.

So, what they are asking? Population mean is given, population variance is given, we have to find out the range of sample means that is \bar{X} , lower limit, upper limit. By the central limit theorem, we know that the distribution of X is approximately normal, if n is large enough with a mean μ and standard deviation. Let $Z_{\alpha/2}$ be the Z value that leaves area $\alpha/2$, in the upper tail of the normal distribution, that is in the interval $\pm Z_{\alpha/2}$, encloses probability $(1 - \alpha)$.

$$\mu + z_{\alpha/2} \sigma_{\bar{X}}$$

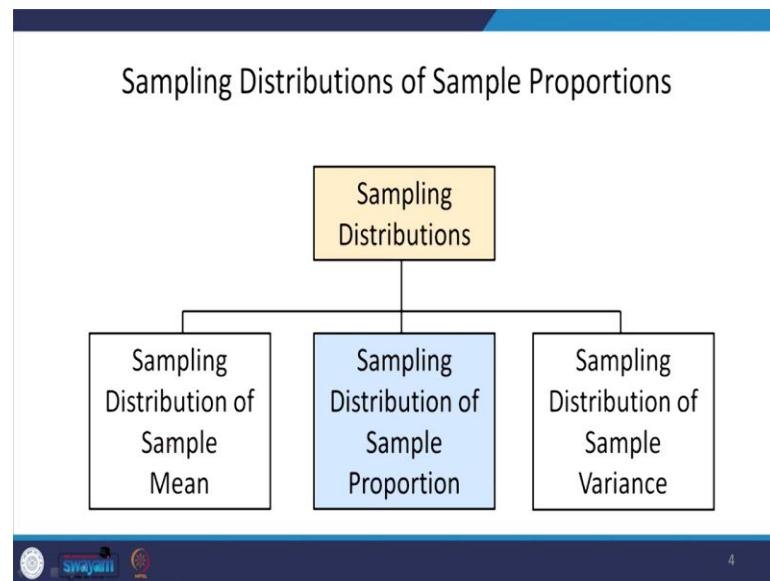
So we can find out the upper limit, the upper limit is .

$$\mu - z_{\alpha/2} \sigma_{\bar{X}}$$

The lower limit is , actually this has come from this formula very, very famous \bar{X} bar minus μ by $\sigma_{\bar{X}}$. From this relationship we can say μ if you re-adjust that you will get this equation okay. So, if you get you know you from this you can find out the \bar{X} bar. So,

this value is X bar value we can get the upper limit and lower limit of X bar is a sample mean okay.

(Refer Slide Time: 05:43)



This was what we have started in the last class. So, sampling distribution there are three things which you are going to see. One is sampling distribution of sample mean which I have seen. This class, we are going to see the sampling distribution of sample proportion and sampling distribution of sample variance. First you will see sampling distribution of sample proportion.

(Refer Slide Time: 06:06)

P = the proportion of the population having some characteristic

- Sample proportion (\hat{p}) provides an estimate of P :
- $0 \leq \hat{p} \leq 1$

$$\hat{p} = \frac{X}{n} = \frac{\text{number of items in the sample having the characteristic of interest}}{\text{sample size}}$$

P equal to the proportion of populations having some characteristics, we can call it as P is the population proportion. This sample proportion we are going to call it as a small \hat{p} . It provides an estimate of P.

(Refer Slide Time: 06:24)

Sampling Distributions of Sample Proportions

P = the proportion of the population having some characteristic

- Sample proportion (\hat{p}) provides an estimate of P:

$$\hat{p} = \frac{X}{n} = \frac{\text{number of items in the sample having the characteristic of interest}}{\text{sample size}}$$

- $0 \leq \hat{p} \leq 1$
- \hat{p} has a binomial distribution, but can be approximated by a normal distribution when $nP(1 - P) > 5$



5

What is the meaning of this estimate of P is sampling distribution of sample proportion. We are going to use capital P, the proportion of population having some characteristics. Then, sample proportion we are going to call it as \hat{p} provides an estimate of capital P. so, what is the meaning of this one is, with the help of sample proportion we can find out the estimate of population proportion.

So, here how the sample proposed is found equal to X divided by n, X is number of items in the samples sample having the characteristics of interest divided by n is sample size the range of sample proportion is as usual zero $\leq \hat{p} \leq 1$. \hat{p} has the binomial distribution, but can be approximated by a normal distribution when $n P Q$ is greater than 5. Here, Q is nothing but 1 minus P so here it is following binomial distribution.

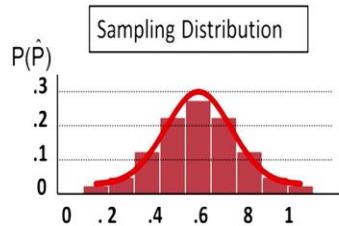
As we know the binomial distribution having properties of having only two alternatives that is good are defective, pass or fail, yes or no. So, only two alternatives is there okay. So what will you do?

(Refer Slide Time: 07:41)

Sampling Distribution of \hat{P}

- Normal approximation:

Properties: $E(\hat{P}) = P$



(where P = population proportion)

And

$$\sigma_{\hat{P}}^2 = \text{Var}\left(\frac{X}{n}\right) = \frac{P(1-P)}{n}$$



6

From the population, we will take sampling proportion so, when you plot the sampling proportion that will follow a normal distribution. So, what will happen? This picture shows the different sample as taken from the population for each sample we find out the sampling proportion, if you plot that sampling proportion that will follow a normal distribution. When we know that it is following normal distribution will have two parameters.

So, mean of that sampling proportion is that is the expected value of your \hat{P} is nothing but P the population proportion. And the variance of this sampling distribution is PQ / n that is a $P(1-P) / n$.

(Refer Slide Time: 08:27)

$$\begin{aligned}
 U_x &= np \\
 \sigma_x^2 &= npq \\
 \sigma_{\hat{P}}^2 &= \frac{\sigma_x^2}{n} = \frac{npq}{n} = pq \\
 U_{\hat{P}} &= \frac{U_x}{n} = \frac{np}{n} = p
 \end{aligned}$$



7

Actually this need not remember this formula we can derive it because we know, we have seen in the previous class, the mean of binomial distribution is nP , the variance of binomial distribution is nPQ . Actually we have to use capital P for the population so we will, I use capital P . Otherwise we can write $1 - P$. Suppose what happen there is a population I am taking proportion 1, proportion 2, proportion 3, proportion 4, like that I may get say P_1 hat, P_2 hat, I will get too many such proportion.

If I plot this, if I plot this sampling proportion that will follow, normal distribution so, we have to find out what is the mean of this sampling proportion distribution. Similarly, what is the variance of this sampling proportion distribution? We know that since we have taken n sample from the

population so the $\mu_{\bar{p}} = \frac{\mu_x}{n}$ okay. We know that not Mu not X bar mu X mu X all know mu X we can write it as $n P / n$ so that is nothing but your population P .

So, what this result says that the mean of the sampling proportion is equal to population proportion. Similarly, according to central limit theorem, this is Sigma by now when you square into Sigma square by n Sigma square by n , so, if you substitute Sigma square is nP I am writing $1 - P$ divided by n this is n square. So, variance is because Sigma square by n .

(Refer Slide Time: 11:01)

Z-Value for Proportions

Standardize \hat{p} to a Z value with the formula:

$$Z = \frac{\hat{p} - P}{\sigma_{\hat{p}}} = \frac{\hat{p} - P}{\sqrt{\frac{P(1-P)}{n}}}$$


8

So, the Z value for the proportion is so small $P \cap$ minus capital P divided by Sigma P we know that Sigma P is root of P into 1 minus P divided by n so P minus so it P minus capital P here

capital P represents the population proportion, \hat{p} represents the sample proportion.

(Refer Slide Time: 11:23)

Example

- If the true proportion of voters who support Proposition A is $P = .4$, what is the probability that a sample of size 200 yields a sample proportion between .40 and .45?
- i.e.:

**if $P = .4$ and $n = 200$, what is
 $P(.40 \leq \hat{p} \leq .45)$?**

9

We do one small problem. If the two proportion of voters who support proposition A is P equal to 0.4, what is the probability that a sample of size 200 yields, a sample proportion between 0.40 to 0.45? What is asked here is the population proportion is given that is a 0.4 that is a 40%. What is the probability that the sample proportion will lie between 0.4 and 0.45.

(Refer Slide Time: 12:03)

Example

(continued)

- if $P = .4$ and $n = 200$, what is
 $P(.40 \leq \hat{p} \leq .45)$?

Find: $\sigma_{\hat{p}}$

$$\sigma_{\hat{p}} = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{.4(1-.4)}{200}} = .03464$$

$$\frac{\hat{p} - P}{\sigma_{\hat{p}}}$$

Convert to standard normal:
normal:

$$P(.40 \leq \hat{p} \leq .45) = P\left(\frac{.40 - .40}{.03464} \leq Z \leq \frac{.45 - .40}{.03464}\right) \\ = P(0 \leq Z \leq 1.44)$$

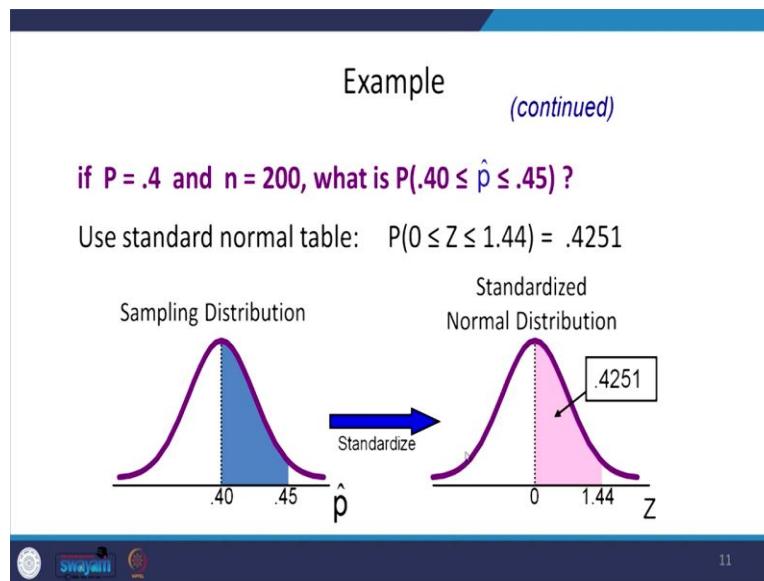
10

So, n is taken equal to 0.4 and n equal to 200 what is the probability of $P(0.4 \leq \hat{p} \leq 0.45)$ here p is sampling proportion less than or equal to 0.45. First you will find out the Sigma proportion that is the standard deviation of sampling proportion. Sigma P equal to root of $P Q$ by n so P is given 0.4, 1 minus 0.42 by 200 we got point 0.34. We have to convert this 1 Sigma P equal to standard normal distribution by using X minus mu / Sigma P so X is given.

X is 0.4 minus so we have find out the standard deviation of sampling proportion that is the 0.3464 so that we will convert into standard normal so that we can refer the table. So, $P(0.4 \leq \hat{p} \leq 0.45)$. $P(0.4)$ this 0.4 is X equal to the small p -0.4 is that capital P divided by 0.03 because that is this your Sigma P there is nothing but did this form is \hat{p} minus capital P divided by Sigma P .

So, P cap that is a lower limit is 0.4, capital P which is given 0.4 divided by Sigma p so this portion will get 0 and right hand side see the another upper limit of P cap 1 is 0.45 minus 0.42 the divided by 0.034 we got the $P(0 \leq Z \leq 1.44)$.

(Refer Slide Time: 13:44)



When you look at the table $P(0 \leq Z \leq 1.44)$. we can get 0.425. So, will summarize what we have done the, it, was asked what is the, what is the probability sampling proportion to lie between 0.4 and 0.45. So, what we are done this 0.4 we are converting to corresponding Z scale it becomes zero. This 0.45, we converted to corresponding Z scale it is 1.44 then we found this area between

Z value is 0 to 1.44 which we got 0.4251. So, now we have seen this one we will go to the sampling distribution of sample variance.

(Refer Slide Time: 14:27)

Sample Variance

- Let x_1, x_2, \dots, x_n be a random sample from a population. The **sample variance** is
$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$
- the square root of the sample variance is called the **sample standard deviation**
- the sample variance is different for different random samples from the same population



13

Let X_1, X_2 and X_n be the random sample from a population, the sample variance is sample

$$\frac{\sum (X_i - \bar{X})^2}{n-1}$$

variance. The square root of the sample variance is called the sample standard deviation. The sample variance is different for different random samples from the same population, because every time you may get different sample variance, okay, very important result which we are going to see.

(Refer Slide Time: 14:55)

Sampling Distribution of Sample Variances

- The sampling distribution of s^2 has mean σ^2

$$E(s^2) = \sigma^2$$

- If the population distribution is normal then

$$\frac{(n-1)s^2}{\sigma^2}$$

has a **χ^2 distribution** with $n - 1$ degrees of freedom



14

The sampling distribution of sample variance has the mean population variance. So, what is the meaning in that one is, from the population, you take different sample for that sample you find the sample variance we know of that sample variance is equal to population variance but when you take the from the normal population, if you take some sample, then, you find the sample variance.

If you plot that that will follow a particular distribution that shape of this will be like this, right skewed distribution. That distribution is called chi-square distribution that you will see in the next slide. So, another important result is if the population distribution is normal then there is a relationship between sample variance and population variance. That is that relation is $(n-1)s^2 / \sigma^2$ as a chi-squared distribution with the n minus 1 degrees of freedom.

So this x axis is nothing but $(n-1)s^2 / \sigma^2$. This is nothing but our chi-square distribution. You may see there is a similarity between, there may be intuitively you can connect with the normal distribution. For example, we say that we will see in the next slide.

(Refer Slide Time: 16:16)

The whiteboard contains the following handwritten content:

$$(Z) = \frac{(x-\mu)^2}{\sigma^2}$$

$$\chi^2 = \frac{\sum(x-\mu)^2}{\sigma^2}$$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

To the right of these formulas is a sketch of a chi-square distribution curve, which is skewed to the right (positively skewed).

For example $Z = (X - \mu)/\sigma$, you take different $X_1 X_2 X_3$ variable so you will get $\Sigma(X - \mu)$. So, what will happen when you square both side when you square both side for different degrees of freedom, so like that you take different sample different means different X_1, X_2, X_3 so this will become $\Sigma(X - \mu)^2/\sigma^2$.

So, the square of Z this is will become a chi square. So, what is nothing but Chi square is nothing but $\sum(X - \mu)^2$. I think you know this formula the variance is $\sum(X - X\bar{ })^2 / (n-1)$, sample variance. So, this numerator can be replaced by that is $\sum(X - X\bar{ })^2$ can be replaced by $(n-1) s^2 / \sigma^2$. That is nothing but your chi-square distribution.

So there is a connection between your Z distribution and chi-square distribution the other thing since it is a squared you see that Z it is normal distribution this way, so chi-square distribution is like this, because you see that we have squared that Z value, so that there will not be negative Z; so Chi square will be always positive. That is the connection between your Z distribution and Chi-square distribution.

(Refer Slide Time: 18:00)

The Chi-square Distribution

- The chi-square distribution is a family of distributions, depending on degrees of freedom: $d.f. = n - 1$

16

What will happen? From the sample, you would have taken the variance when you plot that sample variance that will follow this shape. So, this x axis is nothing but your chi-square value. So, the chi-squared distribution is your family of distribution depending on the degrees of freedom $n-1$. So, when the degrees of freedom it is increasing that means if you are started to take more samples from the population, then, you plot that the variance at the end that will follow a normal distribution.

What will happen? Your chi-square distribution if the degrees of freedom has increased, that will follow a normal distribution. What is the chi-square distribution? From the population, you take some sample, for that sample, you find the variance, like that you take many sample you will find different variance when you plot that variance that will follow this shape. This shape is nothing but the chi-square distribution. What is this chi-square distribution? This x-axis is $(n-1) s^2 / \sigma^2$, okay.

(Refer Slide Time: 19:03)

Degrees of Freedom (df)

Idea: Number of observations that are free to vary after sample mean has been calculated

Example: Suppose the mean of 3 numbers is 8.0

Let $X_1 = 7$
Let $X_2 = 8$
What is X_3 ?

If the mean of these three values is 8.0,
then X_3 must be 9
(i.e., X_3 is not free to vary)

Here, $n = 3$, so degrees of freedom $= n - 1 = 3 - 1 = 2$

(2 values can be any numbers, but the third is not free to vary for a given mean)



17

Then, another important concept is degrees of freedom because many of the time we will use this concept degrees of freedom, we will see, what is the degrees of freedom? Number of observations that are free to vary after a sample mean has been calculated. That is the degrees of freedom. Suppose that the mean of 3 numbers is 8, say 8 so x_1 equal to 7, x_2 equal to 8 what is the value of x_3 what will happen? Since already the mean is known to us we can supply any value to x_1 any value to x_2 .

But you cannot give any value to x_3 because we have lost one degrees of freedom because already we know, what is the mean of that? So what is the logic here is when n equal to 3 so the degrees of freedom is $n - 1 = 2$ values can be any numbers but the third is not free to vary from given mean. It is like, example like, assume that there are three chair is there we are asking three student to sit there. The first person who is entering will have three possibilities.

That is the three degrees of freedom because three chairs are available. The second person will have two possibilities there is a two degrees of freedom. The third one but there is only one chair there is no option for that so you are lost one degrees of freedom. There are if there are n values you will have only n minus 1 degrees of freedom just we have introduced what is the chi-square distribution and how it has connection with the normal distribution. We will do a small problem to understand the application of chi-square distribution.

(Refer Slide Time: 20:34)

Chi-square Example

- A commercial freezer must hold a selected temperature with little variation. Specifications call for a standard deviation of no more than 4 degrees (a variance of 16 degrees²).
- A sample of 14 freezers is to be tested
- What is the upper limit (K) for the sample variance such that the probability of exceeding this limit, given that the population standard deviation is 4, is less than 0.05?

18

A commercial freezer must hold their selected temperature with a little variation specification called for a standard deviation of no more than 4 degrees that is the variance 16 degree square you should not exceed 16, and the standard deviation 4. For a sample of 14 freezers is to be tested what is the upper limit of the sample variance such that the probability of exceeding this limit given that the population standard deviation is 4 is less than 0.05.

What is it asking, what is the probability of sample variance that the, the probability of exceeding this limit is less than 0.05? You will see the next slide what it says exactly.

(Refer Slide Time: 21:25)

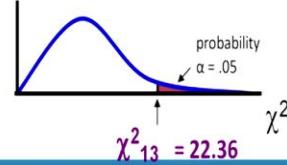
Finding the Chi-square Value

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Is chi-square distributed with $(n - 1) = 13$ degrees of freedom

- Use the chi-square distribution with area 0.05 in the upper tail:

$$\chi^2_{13} = 22.36 \text{ } (\alpha = .05 \text{ and } 14 - 1 = 13 \text{ d.f.})$$



19

So, first thing is we have to find out the Chi square value for n minus 1 degrees of freedom. This is a chi-square distribution there are 14 sample is n the degrees of freedom is for 13. 14 minus 1 13, so, the corresponding alpha is equal to 0.05, is 22.36.

(Refer Slide Time: 21:46)

Chi-square Example

(continued)

$$\chi^2_{13} = 22.36 \text{ } (\alpha = .05 \text{ and } 14 - 1 = 13 \text{ d.f.})$$

So: $P(s^2 > K) = P\left(\frac{(n-1)s^2}{16} > \chi^2_{13}\right) = 0.05$

or $\frac{(n-1)K}{16} = 22.36$ (where $n = 14$)

so $K = \frac{(22.36)(16)}{(14-1)} = 27.52$

If s^2 from the sample of size $n = 14$ is greater than 27.52, there is strong evidence to suggest the population variance exceeds 16.



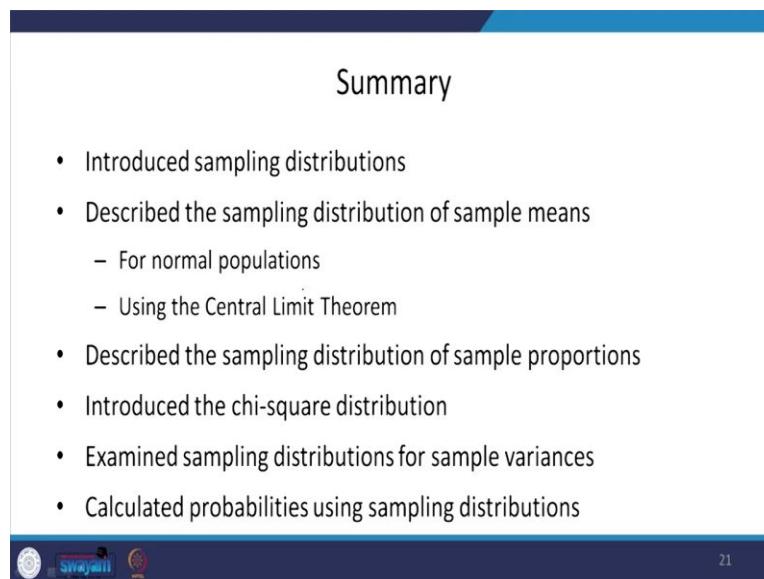
20

So, what is asked is, if, if the chi is 22.36 right we know that the P of $(n-1) s^2$ by, Sigma is 4 so σ^2 is 16 if the chi-square value is 22.36, what is the value of your sample variance that was the question. What is asked the chi-square value is known to us that is 22.36 when alpha equal to 0.05 when chi-square value is 22.36 what is the maximum value of your sample variance.

So, probability of ($s^2 > k$) equal to $\{P(n - 1) s^2 / 16\}$ greater than chi-square 13, equal to 0.05. So, this value this value, this value, $(n - 1) s^2 / 16$ so with this s^2 between that is your K. So, $(n - 1) K / 16$ equal to 22.36 and you simplify we are getting 27.52. The result is give the sample variance from the sample size of 14 is greater than 27.52 there is a strong evidence to suggest that the population variance exceeds 16.

That is the application of this chi-square distribution. We will see in detail there are many applications for a chi-square distribution one is test of Independence, another one is good goodness of it that we will see in coming classes.

(Refer Slide Time: 23:11)



The slide has a blue header bar and a blue footer bar. The main content area is white. The title 'Summary' is centered at the top. Below it is a bulleted list of ten items. At the bottom of the slide, there is a footer bar containing three small icons and the number '21' on the right side.

- Introduced sampling distributions
- Described the sampling distribution of sample means
 - For normal populations
 - Using the Central Limit Theorem
- Described the sampling distribution of sample proportions
- Introduced the chi-square distribution
- Examined sampling distributions for sample variances
- Calculated probabilities using sampling distributions

Now we will summarize in this class what we have seen we have introduced what is the sampling distributions described the sampling distribution of sample means for a normal population. Then we have explained what is the central limit theorem, then, we have seen the sampling distribution of mean, then we have seen the sampling distribution of variance, then, we have seen the sampling distribution of proportions.

Then I have introduced the concept of chi-square distribution how it has connection with normal distribution. Then, we have seen application of chi-square distribution. The next class will go to the next topic, Confidence Interval. We will continue in the next class. Thank you.

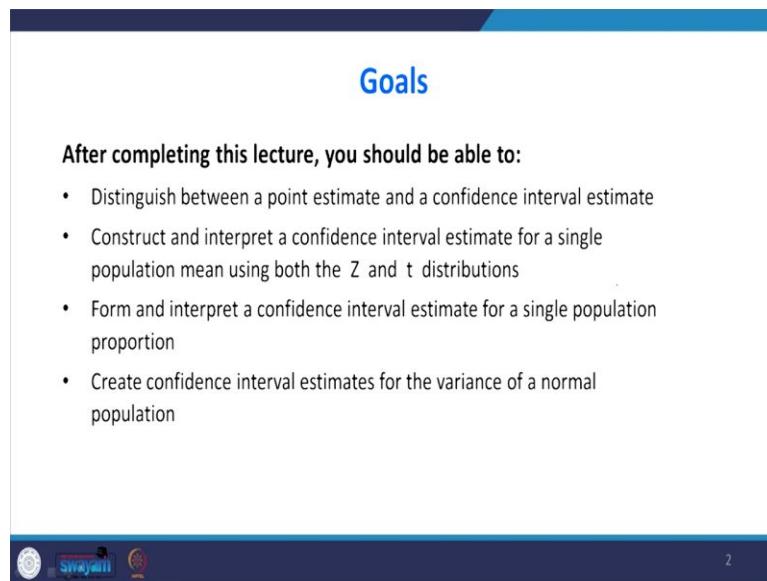
Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 14
Confidence Interval Estimation:
Single Population-1

Welcome students to the next lecture. Today lecture is we are going to talk about the confidence interval estimation for the single population. In the previous lecture, we have seen the sampling distribution. Sampling distribution we have seen three results, out of the sampling distribution lecture. One is sampling distribution for mean, sampling distribution for proportion and sampling distribution for variance.

By using that result we are going to estimate some population parameters. What we are going to estimate? We may estimate the mean of the population or the proportion of the population or the variance of the population. That we will see in this lecture.

(Refer Slide Time: 01:15)



The slide has a dark blue header and footer. The main content area is white with a blue horizontal bar at the top. The title 'Goals' is centered in blue text. Below it is a bulleted list of learning objectives in black text. At the bottom, there is a dark blue footer bar with small white icons and the number '2'.

Goals

After completing this lecture, you should be able to:

- Distinguish between a point estimate and a confidence interval estimate
- Construct and interpret a confidence interval estimate for a single population mean using both the Z and t distributions
- Form and interpret a confidence interval estimate for a single population proportion
- Create confidence interval estimates for the variance of a normal population

The objective of this course is to distinguish between a point estimate and a confidence interval estimate and construct and interpret a confidence interval estimate for a single population mean using both Z and t distributions. Today also in this lecture I am going to introduce Z and t

distribution then, form and interpret, a Confidence Interval estimate for a single population proportion, create confidence interval estimate for the variance of the normal population.

(Refer Slide Time: 01:46)

Confidence Intervals

- Confidence Intervals for the Population Mean, μ
 - when Population Variance σ^2 is Known
 - when Population Variance σ^2 is Unknown
- Confidence Intervals for the Population Proportion, \hat{p} (large samples)
- Confidence interval estimates for the variance of a normal population

3

In the confidence interval, what we are going to see today confidence intervals for the population mean there are two possibilities: When the population variance Sigma square is known, other case is when population variance Sigma square is unknown. Then, confidence interval for the population proportion is P hat, using large samples. Then, confidence interval estimate for the variance of your normal distribution.

(Refer Slide Time: 02:14)

Definitions

- An estimator of a population parameter is
 - a random variable that depends on sample information ...
 - whose value provides an approximation to this unknown parameter
- A specific value of that random variable is called an estimate

$$\bar{x} \rightarrow \mu \quad \hat{p} \rightarrow p$$
$$s^2 \rightarrow \sigma^2$$

4

Before getting into the content, we will see what is the estimator and estimate. An estimate of your population parameter is a random variable that depends on sample information. An estimator whose value provides an approximation to the unknown parameter for example, a specific value of the random variable is called estimate. For example \bar{X} bar is estimator for population mean.

Similarly S^2 , sample variance an estimator for population variance $P\hat{\sigma}^2$, population proportion is estimated for $P\hat{\sigma}$ normal proportion is estimated for population proportion. So, this \bar{X} bar S^2 $P\hat{\sigma}$ these are called estimator.

(Refer Slide Time: 03:16)

Point and Interval Estimates

- A **point estimate** is a single number,
- a **confidence interval** provides additional information about variability

5

A specific value of \bar{X} bar, S^2 , $P\hat{\sigma}^2$ is nothing but estimate. In estimate, there are two things we can say, one is a point estimate is a single number other one is a confidence interval provides additional information about the variability. We can say it is a point estimate and interval estimate because in point estimate is only single number. It is not very much reliable but the confidence interval is giving you additional information about the variability of that point estimate.

For example when you look at this picture, we say what is a point estimate and interval estimate. A point estimate is a single number and interval estimate provides additional information about the variability of the point estimate. For example, you see that if I, if I say tomorrow what is

going to be temperature if I say is exactly 35 degree Celsius, this is a point estimate. If I give some lower limit and upper limit for this for example, this may be say 30 to 40 that is the confidence interval.

And I say 35 it is a single number but when I say 30 to 40 that is in confidence interval. So, the 30 can be called as a lower confidence limit the right side can be called as upper control limit 40. So this is width of confidence interval the point estimate is just one number, single number.

(Refer Slide Time: 04:40)

The slide has a blue header bar with the title "Point Estimates". Below the title is a table with three rows and three columns. The first row has two columns: "We can estimate a Population Parameter ..." and "with a Sample Statistic (a Point Estimate)". The second row has two columns: "Mean" and " μ ". The third row has two columns: "Proportion" and "P". To the right of the table, there is a small orange box containing the symbol " \bar{x} ". Below the table, there is a small orange box containing the symbol " \hat{p} ". At the bottom of the slide, there is a dark blue footer bar with the Swastik logo and the number "6".

We can estimate a Population Parameter ...	with a Sample Statistic (a Point Estimate)
Mean	μ
Proportion	P

Yes, so, point estimate we can estimate population parameter mean μ with the help of sample mean that is \bar{x} . We can estimate population proportion P with help of sample proportion small p.

(Refer Slide Time: 04:55)

Unbiasedness

- A point estimator $\hat{\theta}$ is said to be an **unbiased estimator** of the parameter θ if the expected value, or mean, of the sampling distribution of $\hat{\theta}$ is θ ,

$$E(\hat{\theta}) = \theta \quad \xrightarrow{\text{X}} U$$

- Examples:
 - The sample mean \bar{x} is an unbiased estimator of μ
 - The sample variance s^2 is an unbiased estimator of σ^2
 - The sample proportion \hat{p} is an unbiased estimator of P



7

Then, another important property of this estimator is it should be Unbiasedness. A point estimator $\hat{\theta}$ is said to be an unbiased estimator of the parameter θ if the expected value or mean of the sampling distribution $\hat{\theta}$ is θ . So, then we can say it is unbiased estimator if the expected value of $\hat{\theta}$ equal to θ , then, we can say it is the unbiased estimate.

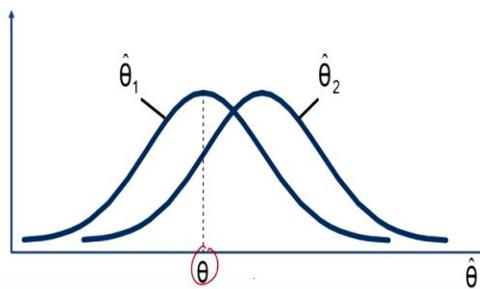
For example if I say \bar{X} then when I can say \bar{X} is an estimator of population proportion. If I say, if the expected value of \bar{X} is equal to mean then we can say \bar{X} is an unbiased estimator, the sample mean \bar{X} is an unbiased estimator for μ the sample variance small s^2 is an unbiased estimator for σ^2 , the sample proportion small p is an unbiased estimator for population proportion P .

(Refer Slide Time: 06:02)

Unbiasedness

(continued)

- $\hat{\theta}_1$ is an unbiased estimator, $\hat{\theta}_2$ is biased:



8

Look at the, another property of unbiasedness. Suppose, if you look at this picture, there are two pictures there. One is for theta one another one is theta two. So, theta one the mean of theta one cap is nothing but your theta. So, the theta one cap is an unbiased estimator. But theta two cap is not unbiased estimator because the mean will be somewhere here, because it is not the population mean, let us see the unbiasedness of an estimator.

There is a two figure is there. One is theta 1 hat another one is theta 2 hat. If you look at the theta 1 cap the mean of theta 1 cap is the theta that is the population mean. But the mean of theta 2 cap is away from the population mean. So, we can say theta 1 cap is an unbiased estimator of the population.

(Refer Slide Time: 06:52)

Bias

- Let $\hat{\theta}$ be an estimator of θ
- The bias in $\hat{\theta}$ is defined as the difference between its mean and θ

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

- The bias of an unbiased estimator is 0

9

We can measure the biasness. Let theta hat be an estimator of theta the bias in theta hat is defined as the difference between the mean and theta. So, the biasness of theta cap is nothing but the expected value of theta hat - theta. The bias of an unbiased estimator is 0 if it is 0 we can say there is no biasness.

(Refer Slide Time: 07:18)

Most Efficient Estimator

- Suppose there are several unbiased estimators of θ
- The most efficient estimator or the minimum variance unbiased estimator of θ is the unbiased estimator with the smallest variance
- Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators of θ , based on the same number of sample observations. Then,

– $\hat{\theta}_1$ is said to be more efficient than $\hat{\theta}_2$ if $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$

– The relative efficiency of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ is the ratio of their variances:

$$\text{Relative Efficiency} = \frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_1)}$$

10

How can we see the most efficient estimator? Suppose there are several unbiased estimator of theta, we have seen sample mean is the one of the estimator of the mean. The most efficient estimator or the minimum variance unbiased estimator of theta is the unbiased estimator with the smallest variance. So, even though there are different estimators to predict the population parameter, we have to see a estimator which is having the smallest variance is the efficient

estimator. Let theta 1 hat and theta 2 hat be the two unbiased estimator of theta. Based on the same number of sample observation then theta 1 hat is said to be more efficient than theta 2 hat, variance of theta 1 hat is less than the variance of theta 2 hat. So what is the point here is there may be different estimator for the population parameter if we want to say which is more efficient we have to see the variance of the estimator.

If the variance of the estimator is lesser then that estimator is the most efficient estimator. The relative efficiency of theta 1 hat with respect to theta 2 is the ratio of their variance. So, relative efficiency is variance of theta 2 hat divided by variance of theta 1 hat.

(Refer Slide Time: 08:41)

Confidence Intervals

- How much uncertainty is associated with a point estimate of a population parameter?
- An interval estimate provides more information about a population characteristic than does a point estimate
- Such interval estimates are called confidence intervals

11

Then, Confidence Interval. How much uncertainty is associated with the point estimate of the population parameter because when I say, the previous example the temperature is 35 degree how much uncertainty is associated with that point estimate. That uncertainty is expressed with the help of confidence interval. An estimate provides more information about the population characteristics than does a point estimate.

So, when compared to point estimate, interval estimate is giving more information about the population. Such interval estimates are called confidence intervals. So, for example, if we say this is the population I am taking different sample say, the population mean may be say 40. I have taken various sample with help of sample mean, I can predict what will be the lower limit

and upper limit of this population mean. For example, if I say, 35 to 45 this interval is nothing but confidence interval. I can go for an exactly endpoint estimate for example if I exactly I can say, point estimate is I can say, exactly say, 40. But the 40 is not much reliable.

(Refer Slide Time: 10:12)

Confidence Interval Estimate

- An interval gives a **range** of values:
 - Takes into consideration variation in sample statistics from sample to sample
 - Based on observation from 1 sample
 - Gives information about closeness to unknown population parameters
 - Stated in terms of level of confidence
 - Can never be 100% confident



12

Confidence interval estimate: An interval give, gives you a range of values. And confidence interval takes into consideration, variation in sample statistics from sample to sample, because what will happen, if there is a big population, we may take different samples but different sample may have different variance, we are constructing the confidence interval with the help of that variance.

So the consideration for the sample to sample is taken with help of, taken into account with help of confidence interval. We can construct the confidence interval based on observation from one sample for example, if we say \bar{X} bar with the help of one sample, I can predict what is the upper limit and lower limit of your μ . It gives information about closeness to unknown population parameters. Stated in terms of level of confidence, can never be 100% confident. We cannot be always 100 % confidence.

(Refer Slide Time: 11:14)

Confidence Interval and Confidence Level

- If $P(a < \theta < b) = 1 - \alpha$ then the interval from a to b is called a $100(1 - \alpha)\%$ confidence interval of θ .
- The quantity $(1 - \alpha)$ is called the confidence level of the interval (α between 0 and 1)
 - In repeated samples of the population, the true value of the parameter θ would be contained in $100(1 - \alpha)\%$ of intervals calculated this way.
 - The confidence interval calculated in this manner is written as $a < \theta < b$ with $100(1 - \alpha)\%$ confidence



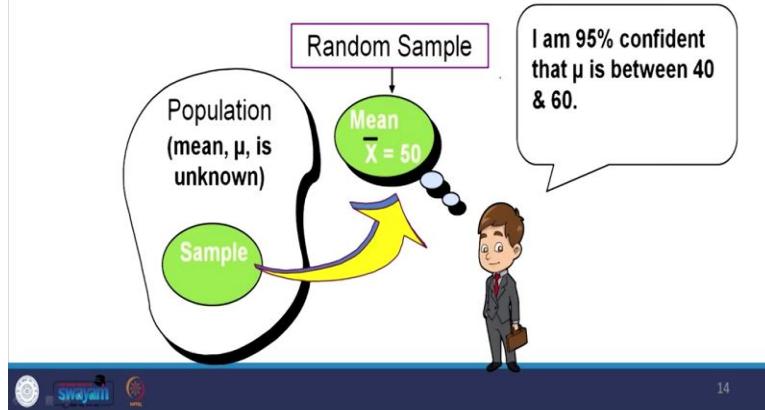
13

Let us see what is the confidence interval and confidence level? So, confidence interval is lower limit and upper limit, the confidence level is nothing but the probability. If $P(a < \theta < b) = 1 - \alpha$ then the interval from a to b is called $100(1 - \alpha)\%$ confidence interval. So, this interval a to b is taken as the confidence interval. The quantity $1 - \alpha$ is called a confidence level. So, confidence level is a probability confidence interval is the lower limit upper limit of population proportion.

So the confidence level alpha is between 0 & 1. In any repeated samples of population the true value of the parameter theta would be contained in $100(1 - \alpha)\%$ of intervals calculated this way. The confidence interval calculated in this manner is written as $a < \theta < b$ with $100(1 - \alpha)\%$ confidence level.

(Refer Slide Time: 12:22)

Estimation Process



14

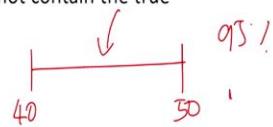
Next we will see what is the estimation process? Look at this the left-hand side this is the whole population mean μ is unknown. We want to predict what is the value of the mean. So, you take the sample the green one say the sample mean is 50, with the help of the sample mean you can say what is the lower limit and upper limit of this population parameter μ , with a certain level of confidence. Say I am saying I am 95% is confident that the μ is between 40 and 60.

(Refer Slide Time: 13:01)

Confidence Level, $(1-\alpha)$

(continued)

- Suppose confidence level = 95%
- Also written $(1 - \alpha) = 0.95$
- A relative frequency interpretation:
 - From repeated samples, 95% of all the confidence intervals that can be constructed will contain the unknown true parameter
- A specific interval either will contain or will not contain the true parameter



15

Then we go to what is a confidence level suppose confidence level is 95 also written as $1 - \alpha$. We will see in detail what is α . α is called a type 1 error. So, $1 - \alpha$ is 0.95, a relative frequency interpretation from a repeated samples 95% of all the confidence intervals that can be

constructed will contain the unknown true parameter. So, what is the meaning of this 95% is, even though we will see in the coming slides.

Suppose if you construct an interval with some range say 40 to 50. So what is the meaning of this 95% so, this interval when you repeat this experiment 100 times, there is a 95% of time you can capture the true mean within this interval. Only 5% of the time this true mean may be outside the interval okay. A specific interval either will contain or will not contain the true parameter.

For example, this interval sometime may contain true parameter otherwise may not contain the true parameter. But when is a 95 % 95 % of the time this interval can't the true parameter there is only five % turns this interval will not capture the true parameter.

(Refer Slide Time: 14:23)

The slide has a blue header bar. The title 'General Formula' is centered in blue text. Below the title is a bulleted list:

- The general formula for all confidence intervals is:

$\text{Point Estimate} \pm (\text{Reliability Factor})(\text{Standard Error})$

$\bar{x} \pm Z \frac{\sigma}{\sqrt{n}}$

- The value of the reliability factor depends on the desired level of confidence

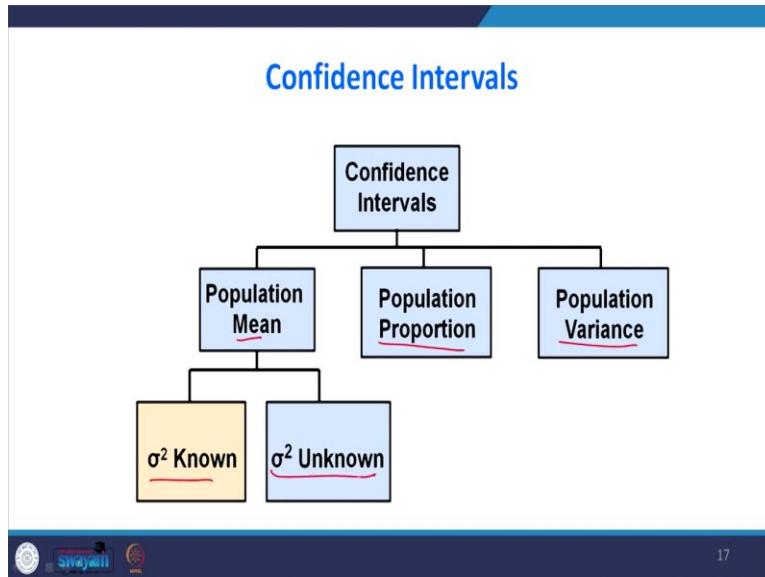
$Z =$

16

The general formula for confidence interval is point estimate is, general formula for point estimate is nothing but your \bar{x} + or - this reliability factor we will see later, Z . This is standard error. If you use a standard error, σ/\sqrt{n} , so \bar{x} + or - $Z(\sigma/\sqrt{n})$ is nothing but the formula for confidence interval. So, when you say + it is upper limit if it is - it is lower limit.

The value of the reliability factor depends upon the desired level of confidence. The value of Z is depending upon how much confidence level you need to have that we will see.

(Refer Slide Time: 15:09)



So, the confidence intervals we will see the classification. We can find the confidence interval for the population mean, we can find confidence interval for the population proportion, we can find the confidence interval for the population variance. In this population mean, there are two category one is Sigma square that is the population variance is known, Sigma square is unknown, the population variance is unknown, whenever the Sigma square, whenever there is a capital letter that represents about the population; whenever there is a small letter that represents about the sample.

(Refer Slide Time: 15:41)

Confidence Interval for μ (σ^2 Known)

- Assumptions
 - Population variance σ^2 is known
 - Population is normally distributed
 - If population is not normal, use large sample
- Confidence interval estimate:

$$\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

(where $z_{\alpha/2}$ is the normal distribution value for a probability of $\alpha/2$ in each tail)

$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

18

We will see first one confidence interval for μ . That means we are going to find the confidence interval of population mean. First case is the Sigma square is known. Sigma square is population

standard deviation is known. What assumptions? Population variance Sigma square is known. Population is normally distributed. If the population is not normal, we have to go for large sample size use a large sample.

So, the confidence interval estimate is $\bar{X} - Z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) < \mu < \bar{X} + Z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$, where σ , $\alpha/2$ is the normal distribution value. So, this is nothing but it will be like this, right. So, this one has come from this formula $(\bar{X} - \mu) / \left(\frac{\sigma}{\sqrt{n}} \right)$. When you when you re-adjust this equation you can find the μ upper limit. This is upper limit, lower limit okay.

When you re-adjust this, $Z_{\alpha/2}$ is nothing but because we are finding both the sides, so this value is $\alpha/2$. This value is $\alpha/2$. So, the remaining places that is $1 - \alpha$. So, this $1 - \alpha$ is called Confidence interval. We will say one more term called margin of error.

(Refer Slide Time: 17:19)

Margin of Error

- The confidence interval,

$$\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Can also be written as $\bar{X} \pm ME$
where ME is called the margin of error

$$ME = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$\sigma - \text{Error}$
 $\frac{\sigma}{\sqrt{n}} - \text{Std Error}$
 $z_{\alpha/2} \frac{\sigma}{\sqrt{n}} - ME$

19

The Confidence interval $\bar{X} - z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) < \mu < \bar{X} + z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$, can be written as $\bar{X} +$ or $- ME$. This ME is nothing but margin of error. So, this term, so this term we can call it as margin of error. You should be very careful when we say, error; generally, another name for standard deviation is the error. Therefore if we write Sigma by root n that is standard error if we say $Z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$, that is a margin of error.

All our error, error is nothing but the variation. So, this is error this is we can say this is error. This is standard error, this is margin of error okay. The standard error whenever you go for sampling that Sigma has to be divided by root of n. This is the result of central limit theorem okay.

(Refer Slide Time: 18:28)

Reducing the Margin of Error

$$ME = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

The margin of error can be reduced if

- the population standard deviation can be reduced ($\sigma \downarrow$)
- The sample size is increased ($n \uparrow$)
- The confidence level is decreased, $(1 - \alpha) \downarrow$

✓ ✓ ✓

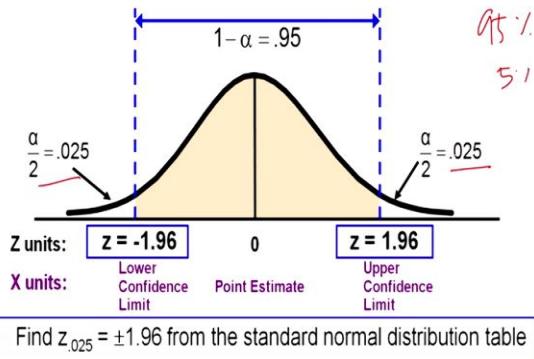
Generally, we have to look for reducing the margin of error. The margin of error can be reduced by looking at this Sigma, n and Z. If the population standard deviation can be reduced when you reduce Sigma, obviously, margin of error will reduce. When you increase the sample size we can predict more accurately the error can be minimized. So, the margin of error will be minimized okay. What is the meaning of this one is, suppose this is one confidence level, this is another confidence level.

For this margin of error for this one, margin of error is more this one, margin of error is more for this one, the margin of error is more. What do I mean whenever the confidence level is small, the margin of error also reduced.

(Refer Slide Time: 19:21)

Finding the Reliability Factor, $z_{\alpha/2}$

- Consider a 95% confidence interval:



21

Then, we look at how to find out the reliability factor that is $Z_{\alpha/2}$. For example, if I suppose, if you want to know something at 95% confidence level, so this is 95% confidence level so the remaining is 5%, when you divide this 5% by 2 see the right hand side you will get is 0.025. The left hand side you will get 0.025. When you look at the Z table, when the right hand side is 0.025, the corresponding Z value is 1.96 on right hand side.

The left side it is - 1.96. This z 1.96 is called upper confidence limit. The left hand side it is called lower confidence limit. The value of Z has to be captured by looking at what is the alpha value. So, when you look at the table the Z value, for 0.025 is + or - 1.96 is from the standard normal table. This we can find out.

(Refer Slide Time: 20:25)

Common Levels of Confidence

- Commonly used confidence levels are 90%, 95%, and 99%

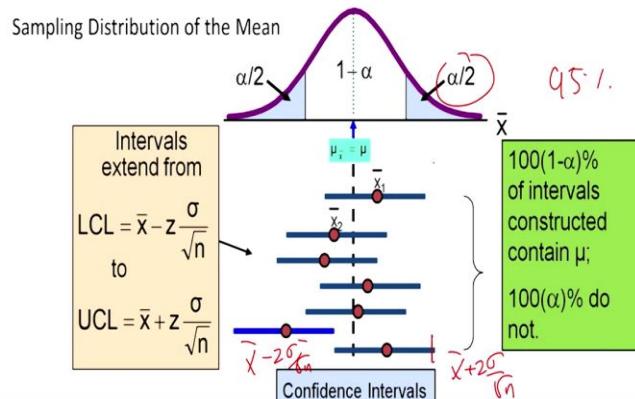
Confidence Level	Confidence Coefficient, $1-\alpha$	$Z_{\alpha/2}$ value
80%	.80	1.28
90%	.90	1.645 ✓
95%	.95	1.96 ✓
98%	.98	2.33
99%	.99	2.58 ✓
99.8%	.998	3.08
99.9%	.999	3.27

22

Look at this, suppose, if the confidence level is 80%. This is nothing but $1 - \alpha$. When you look at the table it is 1.28. But it is at 90%, 0.90 when you look at the table it is 1.645. Generally we will go for 90, 95, 99. So, this value can be remembered. Most of the time, we will go for 95, if it is 95, the Z value is 1.96. Z alpha by 2 not exactly Z, it is Z alpha by 2 when it is 99 then the confidence coefficient - alpha is 0.99 then Z alpha by 2 value is 2.58.

(Refer Slide Time: 21:12)

Intervals and Level of Confidence



23

Next we will see intervals and level of confidence. As I told you, you see that so I have captured 7 intervals. Out of 7, one interval is not lying you are not able to capture the blue one. We are not able to capture the true population parameter okay. So, this is nothing but confidence interval.

So, this portion is nothing but your confidence level. So, $100(1 - \alpha)\%$ of intervals constructed contain μ , that is 100α do not.

Interval extended from lower control limit is $\bar{X} - Z(\sigma/\sqrt{n})$, upper control limit is $\bar{X} + Z(\sigma/\sqrt{n})$. This we can say this is nothing but your $\bar{X} + Z(\sigma/\sqrt{n})$. This left hand side is $\bar{X} - Z(\sigma/\sqrt{n})$. If I say 95% level of confidence what is the meaning is, if I constructed it 100 times, out of 100 times, 95 time my interval which I have constructed will capture the true population mean. Only 5% of time it may not capture the true population parameter.

(Refer Slide Time: 22:37)

Example

- A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms.
- Determine a 95% confidence interval for the true mean resistance of the population.

24

Example a sample of 11 circuits from your large normal population has a mean resistance of 2.20 ohms. Here the sample value is given. This is your sample mean is given. We know from the past testing that the population standard deviation is 0.35 ohms. Determine 95% confidence interval for the true mean resistance of the population.

(Refer Slide Time: 23:06)

Example

(continued)

- A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is .35 ohms.

- Solution:

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

$$= 2.20 \pm 1.96 (.35/\sqrt{11})$$

$$= 2.20 \pm .2068$$

$$1.9932 < \mu < 2.4068$$



25

So, what is given is this is n, this is your 2.20 the sample mean. So, \bar{x} is given 2.20, + Z because the Z value which we got from the table because it is a 95% confidence level. But it is a 95 % confidence level then, the Z value is 1.96 Sigma value is 0.35 is given. There are 11 samples root of n. so, when I say this one the lower limit is 1.9932, the upper limit value is 2.4068.

(Refer Slide Time: 23:39)

Interpretation

- We are 95% confident that the true mean resistance is between 1.9932 and 2.4068 ohms
- Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean

5% \rightarrow Significance Level \rightarrow Type-I-Error
 \rightarrow Producer's Risk

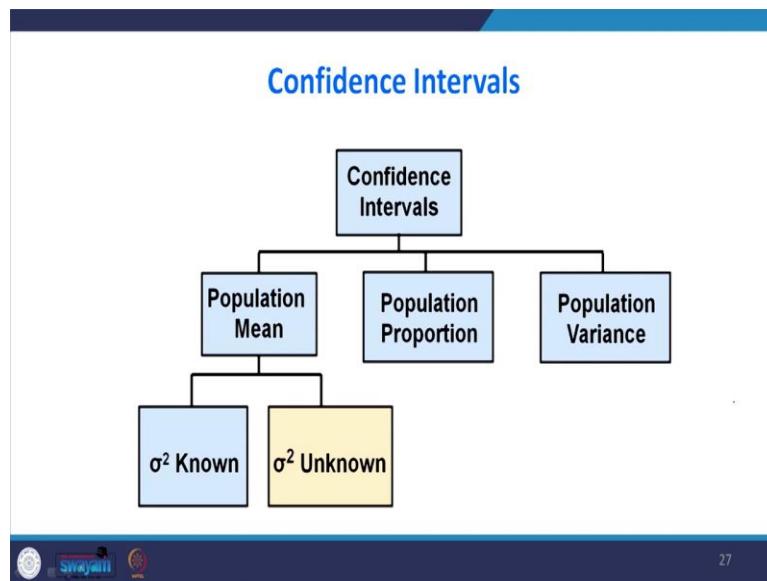


26

So, how we are to interpret this is, we are 95% confident that the true mean resistance is between 1.9932 and 2.4068 ohms. Although the true mean may or may not mean in this interval. 95 % of intervals formed in this manner will contain true mean. Only 5% of time this may not have the

true mean. That is called your significance level. Another name is called type 1 error. Another name is called producers risk. This will see in detail in coming lectures ok.

(Refer Slide Time: 24:35)



We will go to the next category. We will predict the confidence interval or the mean when Sigma square that is the population variance is not given. Dear students I will summarize what we have done so far. We have seen what is the point estimate; we have seen what is the interval estimate we have seen advantage of interval estimate then, we have seen what is the meaning of confidence level.

Then confidence interval after that we have seen how to predict the confidence interval of a population mean when Sigma square is known. In the next lecture will go for predicting the population mean when Sigma square is unknown, thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 15
Confidence Interval Estimation_ Single Population – II

Dear students in the previous class we have predicted the population mean with the help of sample mean where the condition was the Sigma square is unknown now No, the Sigma square is known. Now we will see the next case where Sigma square is unknown then we will see how to predict the population mean.

(Refer Slide Time: 00:49)

Student's t Distribution

- Consider a random sample of n observations
 - with mean \bar{x} and standard deviation s
 - from a normally distributed population with mean μ
- Then the variable $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

follows the Student's t distribution with $(n - 1)$ degrees of freedom

For this purpose you have to use student's t distributions consider a random sample of n observations with the mean $x\bar{x}$ and standard deviation yes from a normally distributed population with the mean μ then the variable t is nothing but $(X \bar{x} - \mu) / (s / \sqrt{n})$ you see that there is a connection between Z , Z to be used to write $(X \bar{x} - \mu) / (\sigma / \sqrt{n})$. But in the t distribution what is happening the Sigma is unknown so we are going to use sample standard deviation.

The other thing is this n should be the smaller number it is less than 30. So, when will you go for

t distribution when Sigma is unknown when n is less than 30 then the variable $t = (\bar{X} - \mu) / (s / \sqrt{n})$ follows the student distributions with $n - 1$ degrees of freedom.

(Refer Slide Time: 01:49)

Confidence Interval for μ (σ^2 Unknown)

- If the population standard deviation σ is unknown, we can substitute the sample standard deviation, s
- This introduces extra uncertainty, since s is variable from sample to sample
- So we use the t distribution instead of the normal distribution

29

Now we will see how to predict the conference interval for μ when Sigma Square is unknown if the population standard deviation Sigma is unknown we can substitute the sample standard deviation s , this introduces extra uncertainty since s is variable from sample to sample. So, we use the t distribution instead of the normal distribution.

(Refer Slide Time: 02:16)

Confidence Interval for μ (σ Unknown)

(continued)

- Assumptions
 - Population standard deviation is unknown
 - Population is normally distributed
 - If population is not normal, use large sample
- Use Student's t Distribution
- Confidence Interval Estimate:

$$\bar{X} - t_{n-1, \alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1, \alpha/2} \frac{S}{\sqrt{n}}$$

where $t_{n-1, \alpha/2}$ is the critical value of the t distribution with $n-1$ d.f. and an area of $\alpha/2$ in each tail

30

What is the assumption for the t distribution population standard deviation is unknown population is normally distributed with the population is not normal use very large sample the

student t distribution the confidence interval is $(\bar{x} - (t_{n-1,\alpha/2} \cdot S/\sqrt{n})) < \mu < (\bar{x} + (t_{n-1,\alpha/2} \cdot S/\sqrt{n}))$ so this also came from this $t_{n-1,\alpha/2}$ this has come from this expression. So, when you readjust that this equations then we can get the lower limit upper limit for the population mean.

(Refer Slide Time: 03:07)

Margin of Error

- The confidence interval,

$$\bar{x} - t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}$$

- Can also be written as $\bar{x} \pm ME$

where ME is called the margin of error:

$$ME = t_{n-1,\alpha/2} \frac{\sigma}{\sqrt{n}}$$

31

So, if it is $-$, $(\bar{x} - (t_{n-1,\alpha/2} \cdot S/\sqrt{n}))$ it is a lower limit if it is $+$, $(\bar{x} + (t_{n-1,\alpha/2} \cdot S/\sqrt{n}))$ it is a upper limit. So, this can be written as like our previous Z distribution $X \bar{x} +$ or $- ME$ this margin of error so this mu is nothing but $t \sigma/\sqrt{n}$ previously it was $Z \sigma/\sqrt{n}$ now it is $t \sigma/\sqrt{n}$.

(Refer Slide Time: 03:41)

Student's t Distribution

- The t is a family of distributions
- The t value depends on degrees of freedom (d.f.)
 - Number of observations that are free to vary after sample mean has been calculated

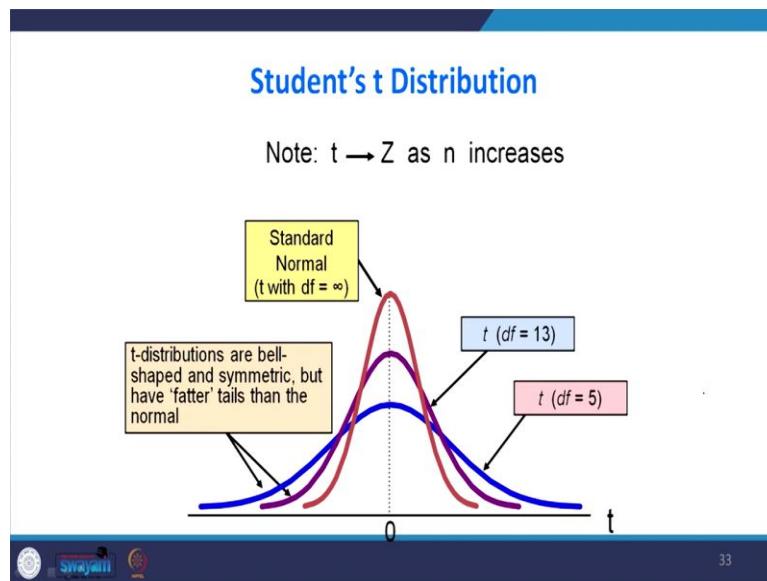
$$d.f. = n - 1$$

32

Student's t-distribution the t is a family of distributions because for every degrees of freedom you will get you a different t distribution. The t value depends upon degrees of freedom number of

observations that are free to vary after sample mean has been calculated nothing but degrees of freedom that is your $n - 1$.

(Refer Slide Time: 04:04)

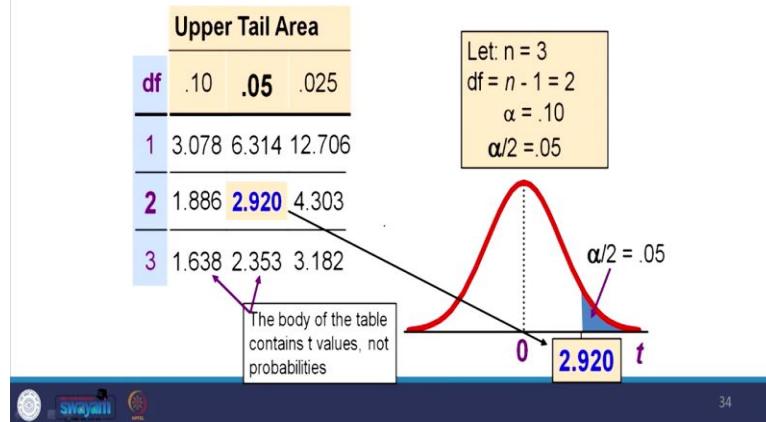


Look at this connection between t distribution and Z distribution, we start from the, see the flatter one t distributions are bell shape and symmetric but have flatter tails than the normal. So, when the degrees of freedom see initially 5 it is a flatter when the data of freedom is 13 and so on you see when the degrees of freedom is infinity it is behaving like a Z distribution that is why in many software packages see there would not be any option for doing Z test there will be option only for doing t test.

Because when the sample size increases for the t test so the behavior of Z distribution t distribution is same.

(Refer Slide Time: 04:45)

Student's t Table



Then we look at the students of t table you said there is a difference between Z table and t table in Z table whatever value which is given is the area but in a t table you see that the area is given on the top say 0.05 the whatever value which is given inside the t table is that is a critical value. For example if it is alpha by 2 is a 0.05 the corresponding t value is 2.920 so the body of the table contains t value not probabilities we should be very careful.

So, for example n equal to 3 and degrees of freedom is $n - 1 = 2$, α equal to 10% then $\alpha/2 = 0.05$ so we are to see where the .05 in the column, column line when degree of freedom is 2 then we can see that is a 2.920.

(Refer Slide Time: 05:41)

t distribution values

With comparison to the Z value

Confidence Level	t (10 d.f.)	t (20 d.f.)	t (30 d.f.)	Z
.80	1.372	1.325	1.310	1.282
.90	1.812	1.725	1.697	1.645
.95	2.228	2.086	2.042	1.960
.99	3.169	2.845	2.750	2.576

Note: $t \rightarrow Z$ as n increases

35

A kind of a comparison between t values and Z values first we will go for this one, this is so familiar for us when the confidence level is 95% see the Z value is 1.96 for different degrees of freedom you see that you see that when the degrees of freedom is 10 it is a 2.228 when is it 20 it is 2.086 see that when t equal to 30 the degrees of freedom is 2.0, so the value of t approaches Z when n increases you see that initially it is increasing so it is starting you know it is decreasing and finally it reaches 1.96.

This table explains whenever the degrees of freedom is increases we are getting Z is close to 1.96 for the t distribution.

(Refer Slide Time: 06:44)

Example

A random sample of $n = 25$ has $\bar{x} = 50$ and $s = 8$. Form a 95% confidence interval for μ

– d.f. = $n - 1 = 24$, so $t_{n-1, \alpha/2} = t_{24, 0.025} = 2.0639$

The confidence interval is $\bar{x} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}$

$$50 - (2.0639) \frac{8}{\sqrt{25}} < \mu < 50 + (2.0639) \frac{8}{\sqrt{25}}$$

$$46.698 < \mu < 53.302$$

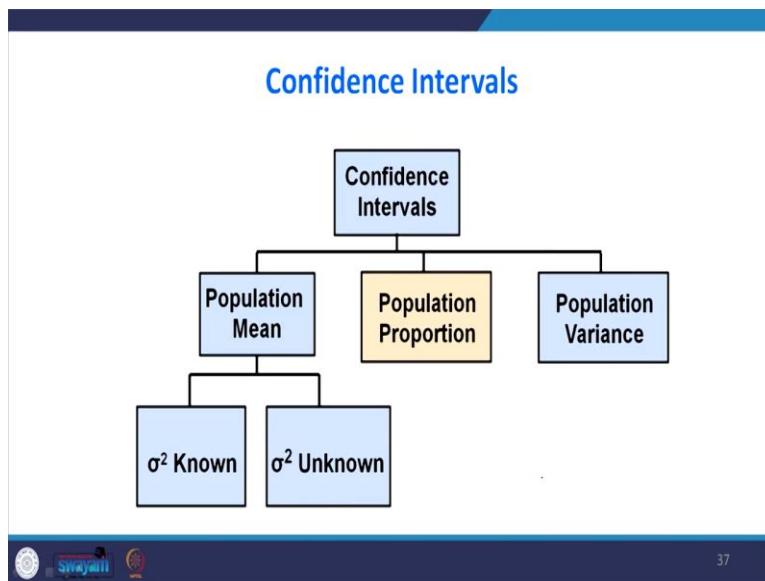
Swastik
www.swastik.com

36

Now we will see how to find out a confidence interval for your t distribution an example is a random sample of n equal to 25 as sample mean is 50 and sample standard deviation is 8 for me a 95% confidence interval for μ . The first one is we have to go for degrees of freedom there are 25 so $25 - 1$, 24 here confidence level is 95% so the significance level is 5% when they say it is a 5% because it is the upper limit lower limit we have to divide by 2 it is 2.5% when degrees of freedom is 24 alpha by 2 is 0.025.

When you look at the table the t value is 2.06 so you substitute X bar equal to 50, t equal to 2.06, S is 8 and sample size is 25 so you are getting lower limit of forty 6.698 upper limit our upper limit of 53.302.

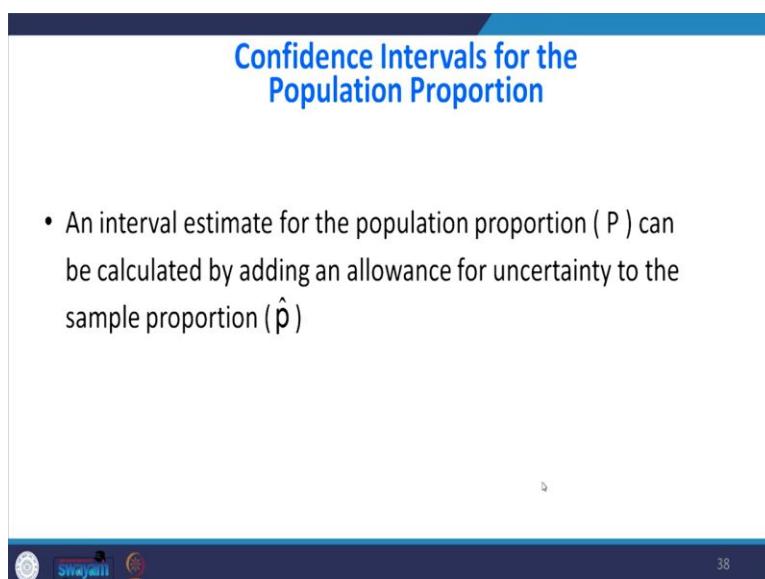
(Refer Slide Time: 07:51)



37

I will go to the next category finding the population proportion with the help of sample proportion.

(Refer Slide Time: 08:04)



38

Confidence interval for the population proportion an interval estimate for the population proportion p can be calculated by adding and elements for uncertain uncertainty to the sample proportion that allowance is nothing but your standard error.

(Refer Slide Time: 08:21)

Confidence Intervals for the Population Proportion, p

(continued)

- Recall that the distribution of the sample proportion is approximately normal if the sample size is large, with standard deviation

$$\sigma_p = \sqrt{\frac{P(1-P)}{n}}$$

- We will estimate this with sample data:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$



39

Recall that the distribution of the sample proportion is approximately normal if the sample size is large then standard deviation is your σ_p , σ_p is root of $P Q$ by n , Q is nothing but $1 - P$ we will estimate this with the sample data. So, this is your sample standard deviation we can say standard

deviation for sampling proportion root of $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

(Refer Slide Time: 08:50)

Confidence Interval Endpoints

- Upper and lower confidence limits for the population proportion are calculated with the formula

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < P < \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- where
 - $z_{\alpha/2}$ is the standard normal value for the level of confidence desired
 - \hat{p} is the sample proportion
 - n is the sample size
 - $nP(1-P) > 5$



40

To find out the lower limit upper limit of the population proportion we have to use the sample values because what will happen we may not know the population P value directly. If you know

population P value what is the purpose of finding lower limit upper limit we know only the

sample proportion so $\hat{P} - z\alpha/2 \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} < P < \hat{P} + z\alpha/2 \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$.

So what is happening so with the help of our sampling proportions we can find out this value is a lower limit of our population proportion. This value is your upper limit of our sampling proportion you see that we with the help of sampling proportion very able to predict. There was a condition but the nPQ should be greater than 5 then only it can be approximated to normal distribution also.

(Refer Slide Time: 09:52)

Example

- A random sample of 100 people shows that 25 are left-handed.
- Form a 95% confidence interval for the true proportion of left-handers

41

An example a random sample of 100 people shows that 25 are left-handed for me a 95% confidence interval for the true proportion of left-handers, so this problem the P cap is 25 by 100 Z is 1.96 because 95% is confidence level all other P cap is given just you substitute this value and then you put plus decide minus you are getting the lower limit of population proportion is 0.1651. The upper limit of population proportion is 0.3349.

(Refer Slide Time: 10:25)

Interpretation

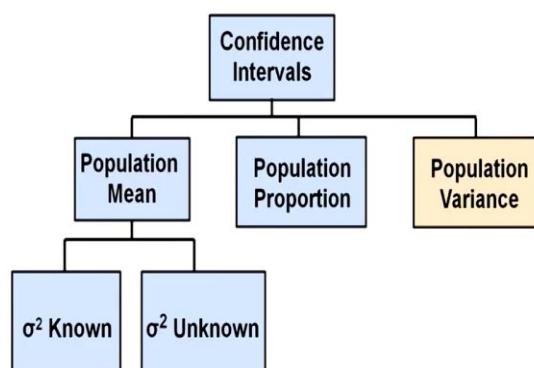
- We are 95% confident that the true percentage of left-handers in the population is between 16.51% and 33.49%.
- Although the interval from 0.1651 to 0.3349 may or may not contain the true proportion, 95% of intervals formed from samples of size 100 in this manner will contain the true proportion.

43

How to interpret this we are 95% confident that the 2% of left-handers in the population is between 16.51% and thirty 33.49% although the interval from 0.1651 to 0.3349 may or may not contain the true proportion 95% of intervals formed from the samples of size 100 in this manner will contain that is more important term. Another way you can say when you repeat this 100 times 95 times you can capture the true population proportion only 5 times you may not capture true population proportion.

(Refer Slide Time: 11:14)

Confidence Intervals



44

We will go to the last one how to predict the population variance. So, so far what we have seen we have predicted the population mean we have predicted the population proportion. Now we are going to predict population variance.

(Refer Slide Time: 11:35)

Confidence Intervals for the Population Variance

- **Goal:** Form a confidence interval for the population variance, σ^2
 - The confidence interval is based on the sample variance, s^2
 - Assumed: the population is normally distributed

45

The goal is to form a confidence interval for the population variance σ^2 . The confidence interval is based on the sample variance. So, what we are going to do with the help of sample variance we are going to predict the population variance interval. We are, assuming the population is normally distributed.

(Refer Slide Time: 11:56)

Confidence Intervals for the Population Variance

(continued)

The random variable

$$\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma^2}$$

follows a chi-square distribution with $(n - 1)$ ¹ degrees of freedom

46

We already have seen that whenever there is a population there if you take some sample from there when you plot the when you plot the sample variance that will follow Chi square distribution as I told you previously it will be like this. This will be $(n - 1) s^2 / \sigma^2$. We are going to use this result when you readjust this right when you readjust this so Sigma Square will be less

than or equal to less than or equal to so, this will become Chi Square n - 1 this side will become n - 1 s square Chi square n - 1 here alpha by 2 here 1 - alpha by 2.

So, what is happening when you readjust this equation when you readjust this equation for Sigma square you can find out the upper limit and lower limit of population variance.

(Refer Slide Time: 13:13)

The $(1 - \alpha)\%$ confidence interval for the population variance is

$$\frac{(n-1)s^2}{\chi_{n-1,\alpha/2}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{n-1,1-\alpha/2}^2}$$

Yes the same thing the $1 - \alpha$ percentage confidence interval for the population variance is given by this one. You look at this the left hand side is alpha by 2 because what will happen when you look at the chi square distribution, we have given only the right side area when the right side area is alpha by 2, so what will happen here they will get to a bigger number. Suppose this was this value is over $1 - \alpha$ by 2.

So, here be a bigger number for example say 5 he will be smaller number when you numerator when you divide by bigger number will become smaller value that will become the lower limit of over variance. The numerator when you divide by a smaller value it will become bigger number that will become the upper limit of your population variance.

(Refer Slide Time: 14:07)

Example

You are testing the speed of a batch of computer processors. You collect the following data (in Mhz):

Sample size 17

Sample mean 3004

Sample std dev 74

Assume the population is normal. Determine the 95% confidence interval for σ_x^2



48

We will see you an example you are testing the speed of batch of computer processors you collect the following data, sample sizes 17 sample mean is 3004 sample standard deviation is 74 assume the population is normal determined 95% confidence interval for $\sigma_{\bar{x}}^2$ here $\sigma_{\bar{x}}^2$ is nothing but lower limit upper limit of the sampling variance.

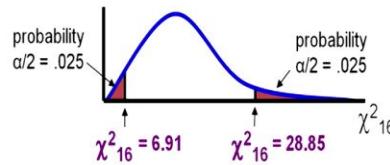
(Refer Slide Time: 14:37)

Finding the Chi-square Values

- $n = 17$ so the chi-square distribution has $(n - 1) = 16$ degrees of freedom
- $\alpha = 0.05$, so use the chi-square values with area 0.025 in each tail:

$$\chi_{n-1, \alpha/2}^2 = \chi_{16, 0.025}^2 = 28.85$$

$$\chi_{n-1, 1-\alpha/2}^2 = \chi_{16, 0.975}^2 = 6.91$$



49

So, n equal to 17 then chi square distribution has the $n - 1$, 16 degrees of freedom when alpha equal to 0.05 because it is we are finding upper limit lower limit we got 2 divided by 2 so 0.025 so when it is alpha by 2 it is 28.25 so what will happen this is the right side limit when you want to know the left side limit you have to, in the chi square table when area equal to 1 - 0.025 that

area you have to find out that probability when the degrees of freedom is 16 so corresponding value is 6.91.

(Refer Slide Time: 15:21)

Calculating the Confidence Limits

- The 95% confidence interval is

$$\frac{(n-1)s^2}{\chi_{n-1, \alpha/2}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{n-1, 1-\alpha/2}^2}$$
$$\frac{(17-1)(74)^2}{28.85} < \sigma^2 < \frac{(17-1)(74)^2}{6.91}$$
$$3037 < \sigma^2 < 12683$$

Converting to standard deviation, we are 95% confident that the population standard deviation of CPU speed is between 55.1 and 112.6 Mhz



50

So when you substitute this value $17 - 1$, s^2 is 74 so this value is chi square value when it is alpha by 2 chi-square value if it is $1 - \alpha$ by 2 you are finding the lower limit is 3037 and upper limit is 12 683 converting the standard deviation we are 95% confident that the population standard deviation of CPU speed is between when you take square root of this between 55.1 and 112.6.

(Refer Slide Time: 15:55)

Finite Populations

- If the sample size is more than 5% of the population size (and sampling is without replacement) then a **finite population correction factor** must be used when calculating the standard error



51

So, far we have assumed that the infinite population sometime there is a finite population. Finite population is when the when the when the number of element in the population is small, if the sample size is more than 5% of the population size and sampling without replacement then the finite population correction factor must be used in calculating standard error. So, we have to add this correction factor when we go for a finite population.

When the finite population we have? when the sample size is more than 5% and being you go for without replacement.

(Refer Slide Time: 16:34)

Finite Population Correction Factor

- Suppose sampling is **without replacement** and the sample size is large relative to the population size
- Assume the population size is large enough to apply the central limit theorem
- Apply the **finite population correction factor** when estimating the population variance

$$\text{finite population correction factor} = \frac{N-n}{N-1}$$



52

Suppose sampling is without replacement and the sample size is large relative to the population size we should go for finite population correction factor. Assume the population size is large enough to apply the central limit theorem. So, apply the finite population correction factor when estimating the population variance, so, this factor $(N - n) / (n - 1)$. So, N is population size is n is sample size.

(Refer Slide Time: 17:06)

Estimating the Population Mean

- Let a simple random sample of size n be taken from a population of N members with mean μ
- The sample mean is an **unbiased estimator** of the population mean μ
- The **point estimate** is:
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$



53

Let the simple random sample of n be taken from the population of n members with μ the sample mean is unbiased estimator of the population mean μ then the point estimator is $(1/n)\bar{X}$, there is no problem for sample mean when you are going for sample variance we have to add this correction factor that is this correction factor has to be added. If the sample size is more than 5% of the population size and unbiased estimator for the variance of the sample mean is s^2 by n you have to multiply this.

(Refer Slide Time: 17:42)

Finite Populations: Mean

- If the sample size is more than 5% of the population size, an unbiased estimator for the variance of the sample mean is

$$\hat{\sigma}_{\bar{x}}^2 = \frac{s^2}{n} \left(\frac{N-n}{N-1} \right)$$

- So the $100(1-\alpha)\%$ confidence interval for the population mean is

$$\bar{x} - t_{n-1, \alpha/2} \hat{\sigma}_{\bar{x}} < \mu < \bar{x} + t_{n-1, \alpha/2} \hat{\sigma}_{\bar{x}}$$



54

So 100 into 1 - alpha% conference interval for the population mean is this μ , $\hat{\sigma}_{\bar{x}}^2$.

(Refer Slide Time: 17:51)

Estimating the Population Proportion

- Let the true population proportion be P
- Let \hat{P} be the sample proportion from n observations from a simple random sample
- The sample proportion, \hat{P} , is an unbiased estimator of the population proportion, P



55

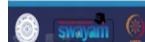
So, this is applicable for population proportion also when the population proportion is population when we are going to predict the population proportion when the sampling proportion is larger and the population proportion is finite then you have to add another correction factor. Let the true population proportion be P , let \hat{P} be the sample proportion from n observation from the simple random sample.

The sample proportion P cap is unbiased estimator of the population proportion n so here also we have to add this $N - n$ by $n - 1$ as a correction factor all others are remaining same.

(Refer Slide Time: 18:37)

Lecture Summary

- Introduced the concept of confidence intervals
- Discussed point estimates
- Developed confidence interval estimates
- Created confidence interval estimates for the mean (σ^2 known)
- Introduced the Student's t distribution
- Determined confidence interval estimates for the mean (σ^2 unknown)



57

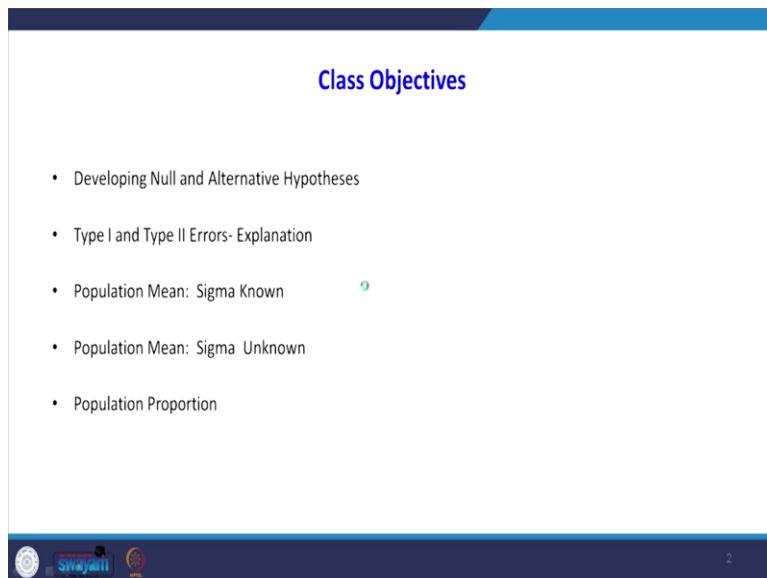
Now we will summarize what we have done so far. In this class we will summarize what we have done so far in this lecture we have created a confidence interval estimate for the proportions then we have created a confidence interval estimate for the variance of a normal distribution. For each proportion and variance estimations we all you taken a numerical example to solve the problem to understand this concept of parameter estimation, thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 16
Hypothesis Testing- I

Welcome students today we are entering to a very, very interesting topic that is on hypothesis testing. Especially this topic is going to be fundamental for coming lectures. So, you are to carefully you should understand.

(Refer Slide Time: 00:40)



Class Objectives

- Developing Null and Alternative Hypotheses
- Type I and Type II Errors- Explanation
- Population Mean: Sigma Known
- Population Mean: Sigma Unknown
- Population Proportion

SWAYAM

The class objectives are first I will to explain how to develop null and alternative hypotheses because solving a hypothesis problem is very easy the most important is how to formulate to the hypothesis. Once you are very good at the formulating the hypothesis solving the problem is very easy. Then I am going to explain what is a type 1 and type 2 error and how this context of type 1 type 2 error is connected with hypothesis.

Next we are going to do hypothesis testing when Sigma that is a population standard deviation is known. Next we will go to hypothesis testing when population standard deviation is not known. Then we will do hypothesis testing for the population proportion.

(Refer Slide Time: 01:27)

Hypothesis Testing

- Hypothesis testing can be used to determine whether a statement about the value of a population parameter should or should not be rejected.
- The null hypothesis, denoted by H_0 , is a tentative assumption about a population parameter
- The alternative hypothesis, denoted by H_a , is the opposite of what is stated in the null hypothesis
- The hypothesis testing procedure uses data from a sample to test the two competing statements indicated by H_0 and H_a .



3

First we will go for what is hypotheses are testing hypothesis testing can be used to determine whether a statement about the value of the population parameter should or should not be rejected. So, hypothesis is nothing but some assumptions about the population parameter we know that most of the populations which we are going to do is going to follow a particular distribution for example a normal distribution.

So normal distribution having two parameter one is mean and variance. So, we can assume the population mean as a hypothesis and assumption otherwise you can have population variance also and hypothesis. The null hypothesis denoted by H_0 is the tentative assumption about the population parameter. So, whatever assumptions which are having that will go to the population parameter the alternative hypothesis denoted by H_a is the opposite of what is stated in the null hypothesis.

The hypothesis testing procedure uses data from the sample to test to computing statement indicated by H_0 and H_a . What are the two competing statement one is null hypothesis another one is alternative hypothesis.

(Refer Slide Time: 02:44)

Developing Null and Alternative Hypotheses

- It is not always obvious how the null and alternative hypotheses should be formulated
- Care must be taken to structure the hypotheses appropriately so that the test conclusion provides the information the researcher wants
- The context of the situation is very important in determining how the hypotheses should be stated
- In some cases it is easier to identify the alternative hypothesis first. In other cases the null is easier
- Correct hypothesis formulation will take practice



4

Next we will see how to develop null and alternative hypothesis, it is not always obvious how the null and alternative hypothesis should be formulated. We should be very careful to structure the hypothesis appropriately so that the test conclusion provides the information the researcher wants. The context of the situation is very important in determining how the hypothesis should be stated. In some cases it is easier to identify the alternative hypothesis first in other cases the null is easier.

So, correct hypothesis formulation, will take a practice in this lecture we are going to take some example and I am going to explain how to formulate the hypothesis whether it is null or altered hypothesis.

(Refer Slide Time: 03:36)

Developing Null and Alternative Hypotheses

Alternative Hypothesis as a Research Hypothesis

- Many applications of hypothesis testing involve an attempt to gather evidence in support of a research hypothesis
- In such cases, it is often best to begin with the alternative hypothesis and make it the conclusion that the researcher hopes to support
- The conclusion that the research hypothesis is true is made if the sample data provide sufficient evidence to show that the null hypothesis can be rejected



5

First you start with alternative hypothesis as a research hypothesis. Most of the time the researchers wanted to prove the alternate hypothesis. Many applications of hypothesis testing involve an attempt to gather evidence in support of research hypothesis. In such cases it is often best to begin with the alternative hypothesis and make it conclusion that the researcher hopes to support it because many of the time the researchers wanted to support his hypothesis.

So first you have to write the alternative hypothesis. The conclusion that the research hypothesis true is made if the sample data provide sufficient evidence to show that the null hypothesis can be rejected so, if we want to accept your alternative hypothesis the data which we are collected from the sample has to support the null hypothesis to reject it.

(Refer Slide Time: 04:36)

Developing Null and Alternative Hypotheses

Alternative Hypothesis as a Research Hypothesis

- Example: A new manufacturing method is believed to be better than the current method.
- Alternative Hypothesis:
 - The new manufacturing method is better
- Null Hypothesis:
 - The new method is no better than the old method



6

Next alternative hypothesis as a research hypothesis example we will see some examples here in example is a new manufacturing method is believed to be better than the current method. Assume that in your manufacturing context some is proposing a new way of doing work a new manufacturing method. So, we want to test this assumption so what is alternative hypothesis the new manufacturing method is better because that new method was given by the researcher always the researcher will believe that whatever he says is there is a support for that.

So, first we will formulate to the alternate hypothesis that is the new manufacturing method is better, the null hypothesis just to the complement of alternate hypothesis so the new method is no better than the old method.

(Refer Slide Time: 05:26)

Developing Null and Alternative Hypotheses

- Alternative Hypothesis as a Research Hypothesis
- Example: A new bonus plan, that is developed in an attempt to increase sales
- Alternative Hypothesis:
 - The new bonus plan increase sales
- Null Hypothesis:
 - The new bonus plan does not increase sales



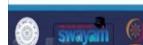
7

We will take another example a new bonus plan that is developed in an attempt to increase the sales. Now what is happening any other organization so we are introducing new bonus plan we are going to see that the bonus has any impact on the sales. So, what is alternative hypothesis is the new bonus plan increases sales. So, first to go for alternate hypothesis then what is the null hypothesis the new bonus plan does not increase the sales you see that whenever when you look at the null hypothesis it will say always does not increase the sales. That is why it is called the null nothing has happened there is a meaning of null.

(Refer Slide Time: 06:13)

Developing Null and Alternative Hypotheses

- Null Hypothesis as an assumption to be challenged
- We might begin with a belief or assumption that a statement about the value of a population parameter is true



9

We will go for another example of alternative hypothesis see a new drug is developed with a goal of lowering cholesterol level more than the existing drug. So, what is happening there are

already there are some drugs available to lower the cholesterol that a researcher has found some drug so that is reducing the cholesterol better than the existing medicine drug.

So, we will go for alternate hypothesis the new drug lowers the cholesterol level more than the existing drug. So, null hypothesis the new drug does not lower the cholesterol level more than the existing drug. You see that that does not so this does not represent the null so nothing significance has happened that is why we are calling it is null hypothesis. Then null hypothesis an assumption to be challenged.

We might begin with the belief or assumption that the statement about the value of the population parameter is true. So, in the hypothesis testing context always we will start the problem assuming that the null hypothesis is true. For example in India before starting a trial suppose somebody was accused so before starting a trial we the trial will be started assuming that the person is innocent person.

You see what is happening the trial will be started assuming that the person is innocent the police has to bring some evidence and they have to say that it is not innocent. When in other countries the person who is being suspected he has to prove his innocence. So, it is reverse so what is the meaning of this reverse that even though something has happened if there is no evidence that person is free.

We then using a hypothesis test to challenge the assumption and determine if there is a statistical evidence to conclude that the assumption is incorrect. In this situation it is helpful to develop the null hypothesis first. We will take an example of how to develop your null hypothesis. A null hypothesis is an assumption to be challenged.

(Refer Slide Time: 08:23)

Developing Null and Alternative Hypotheses

- Null Hypothesis as an Assumption to be Challenged
- Example:
 - The label on a milk bottle states that it contains 1000 ml
- Null Hypothesis:
 - The label is correct. $\mu \geq 1000$ ml



Example you see little the label on your milk bottle states that it contains 1000 ml null hypothesis the label is correct, so $\mu \geq 1000$ ml. Another hint is the null hypothesis is nothing but the status quo in null hypothesis always there will be '=' sign the null hypothesis is looked at a optimistic perspective. If somebody say the bottle contains the 1000 ml we are assuming that yes that assumption is correct so we formulating the null hypothesis is $\mu \geq 1000$ ml.

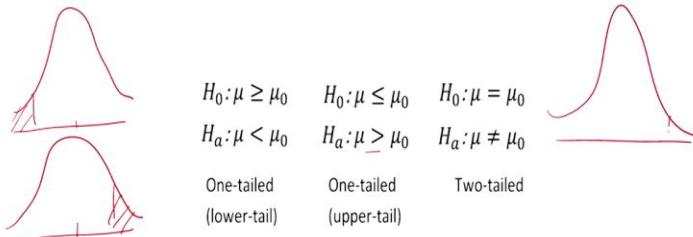
So, alternate hypothesis the label is incorrect $\mu < 1000$ ml you see that the signs are complementary since the null hypothesis it is greater than or equal to there is here less than. If the null hypothesis is less than or equal to so the alternate hypothesis it will be $>$. The null hypothesis is '=' the alternative hypothesis is ' \neq '. And the null hypothesis always will have equal to sign alternate hypothesis never contained equal to sign.

The status quo will go to null hypothesis we have to challenge the status quo that is nothing but your alternate hypothesis.

(Refer Slide Time: 09:49)

Null and Alternative Hypotheses about a Population Mean μ

- The equality part of the hypotheses always appears in the null hypothesis $\underline{\underline{=}}$
- In general, a hypothesis test about the value of a population mean μ must take one of the following three forms (where μ_0 is the hypothesized value of the population mean)



11

Okay you see how the nature of the null hypothesis. The Equality part of the hypothesis always appear in the null hypothesis that means in null hypothesis always there will be a equal to sign. So, when is equal to means it is equivalent to null nothing has happened that is the status quo is maintained as it is. In general the hypothesis test about the value of the population mean μ must we take one of the following 3 forms where μ_0 is the hypothesis value of the population mean.

You see that the hypothesis may take different forms for example μ greater than or equal to μ_0 the μ_0 is what do you have a similar population mean. You see that the null hypothesis there is a greater than or equal to so we are writing in the alternative hypothesis less than because the signs are complementary. So, this test is one tailed test that is called lower tailed test. So, how we are calling it to lower tailed test is for example if I am drawing here.

We have to look at the sign of your alternate hypothesis. The sign of for alternate hypothesis is less than μ_0 so it is left tailed test. If anything goes beyond the left hand side we will reject it. See there is another context μ , $H_0: \mu$ less than or equal to μ_0 so alternate hypothesis is μ greater than μ_0 here also you look at this this is a less than or equal to so complement sign is greater than.

So it is one tailed it is called per tail test look at the sign of our alternate hypothesis it is greater than, so if it is greater than so it is called right tailed test. If anything beyond this point suppose

this mean beyond this point will be rejected the last one is equal to sign μ equal to μ_0 , so $H_a: \mu$ not equal to μ_0 , this is called a two tailed test. So, two tailed test is you see that the rejection area will be on both side so if the value goes below this will reject it the value goes about this rejected what is the meaning of value.

(Refer Slide Time: 12:11)

Null and Alternative Hypotheses

- A major hospital in Chennai provides one of the most comprehensive emergency medical services in the world
- Operating in a multiple hospital system with approximately 10 mobile medical units, the service goal is to respond to medical emergencies with a mean time of 8 minutes or less
- The director of medical services wants to formulate a hypothesis test that could use a sample of emergency response times to determine whether or not the **service goal of 8 minutes or less is being achieved.**



12

I will explain we will take an example to do a hypothesis testing a major hospital in Chennai provides one of the most comprehensive emergency medical services in the world. Operating in here multiple hospital system with approximately 100 mobile medical units that hospital is having 100 it is not 100, it is 10 mobile medical units the service goal is to respond to medical emergencies with a mean time of 8 minutes or less.

So, the problem is that they have 10 mobile medical units they are too whenever there is a emergency they have to respond 8 minutes or less. The director of medical services want to formulate a hypothesis test that could you see a sample of emergency response times to determine whether or not the service goal of the goal of 8 minutes or less is being achieved. Look at this problem the director wanted to test the service goal of 8 minutes or less is being achieved.

See now it is like here alternative hypothesis the researchers wanted to test whether the service goal of 8 minutes or less is achieved. So, what will happen now the status quo the status quo is 8 minutes or less.

(Refer Slide Time: 13:40)

Null and Alternative Hypotheses

$H_0: \mu \leq 8$ The emergency service is meeting the response goal; no follow-up action is necessary.

$H_a: \mu > 8$ The emergency service is not meeting the response goal; appropriate follow-up action is necessary.

where: μ = mean response time for the population of medical emergency requests

Swayam

So what happened the status quo will go to null hypothesis so what is the null hypothesis the emergency service meeting the response goal. So, no follow-up action is required the another name why it is called null hypothesis is when you accept a null hypothesis no follow-up action is required, no course of action is required. So, why we are saying mu less than or equal to 8 because that is the status quo, so, always null hypothesis null hypothesis look at at their optimistic perspective.

So when I say $\mu \leq 8$ you see that the opposite of this what is that the emergency service is not meeting the response goal that is appropriate follow-up action is necessary that is why it is called alternate hypothesis, so mu greater than equal to 8 you see that here it is a less than or equal to 8. So, the sign is complimentary it is greater than equal to 8 while we are writing mu less than or equal to the status go should go to null hypothesis okay where the mu is the mean response time for the population of medical emergency request.

(Refer Slide Time: 14:49)

Type I Error

- Because hypothesis tests are based on sample data, we must allow for the possibility of errors
- A Type I error is rejecting H_0 when it is true *true*
- The probability of making a Type I error when the null hypothesis is called the level of significance $\alpha = 5\%$
- Applications of hypothesis testing that only control the Type I error are often called significance tests



14

So, we will go to what is a type 1 error because hypothesis tests are based on the sample data we must allow for possibility of errors because the conclusion of hypothesis that is to accept or reject is based on sample data. So, always there is a possibility of error. Here type 1 error is rejecting H_0 when it is true, as he told you in the code context somebody is pleading that is innocent but the judge is not accepting his innocent but really is innocent but he was his innocence was rejected that is incorrect rejection, that is a type 1 error.

The probability of making a type 1 error is when the null hypothesis is true when the null hypothesis is called the level of significance. So, level of significance we call it is alpha most of the time it is 5% what is the meaning of this 5% is the probability of incorrectly rejection is only 5%, application of hypothesis testing that one that only control the type one error are often called significance test.

(Refer Slide Time: 16:10)

Type II Error

- A Type II error is accepting H_0 when it is false.
- It is difficult to control for the probability of making a Type II error.
- Statisticians avoid the risk of making a Type II error by using "do not reject H_0 " and not "accept H_0 ".

15

Type 2 error a type two error is accepting H_0 when it is false, it is difficult to control the probability of making a type 2 error status easy and avoid there is a risk of making type 2 error by using do not reject H_0 instead of accept null hypothesis because in the hypothesis context when we concluded we will not say accept null hypothesis we will say do not reject null hypothesis. Because there is no proof for that null hypothesis is true.

(Refer Slide Time: 16:53)

Type I and Type II Errors

		Population Condition	
		H_0 True ($\mu \leq 8$)	H_0 False ($\mu > 8$)
Conclusion	Accept H_0 (Conclude $\mu \leq 8$)	Correct Decision	Type II Error
	Reject H_0 (Conclude $\mu > 8$)	Type I Error	Correct Decision

16

See the context see the population condition is H_0 is true you see that in the conclusion H_0 is true we will see this when you reject H_0 that is called your type 1 error so that is called incorrect rejection. You see the other case the H_0 is false but you have accepted so that is called your type 2 error. So, another name for a type 1 error is incorrect rejection, for type 2 error it is false

acceptance. We can say another example the producer risk we call this alpha, alpha is called type 1 error, beta consumer risk is called type 2 error.

What is the meaning of this producer risk and consumer is case assume that I am the manufacturer I am producer I am producing shaft, so whose diameter is for example the shaft diameter is say 50 mm. Suppose there is a supplier is coming the supplier has taken some sample from my production lot then he is rejected my lot, he says that you were your production level is not meeting our specification that is 50mm.

There is a 2 possibilities there the supplier who has the way he measured is wrong otherwise I made the sample which have kept is not correct. So, that is incorrect rejection even though I have quality good products they have rejected that is an incorrect rejection that is called to produce a risk. So, there is another possibility assume that I am making only 49 mm of shaft again the supplier came he measured is 50, it is 49 but he is measure it is 50 then he has accepted my lot so that is false acceptance that is called a type 2 error.

There are two possibility one is the sample which I have kept that meet all his requirements but my whole lot does not meet meeting his requirement. So, that means my sample is not the representative of the population that is one possibility otherwise the way they have measured it that is wrong, so that is called false acceptance that is a type 2 error. In the next lecture we will see the application of type 2 error in detail.

(Refer Slide Time: 19:29)

p-Value Approach to One-Tailed Hypothesis Testing

- The p -value is the probability, computed using the test statistic, that measures the support (or lack of support) provided by the sample for the null hypothesis
- If the p -value is less than or equal to the level of significance α , the value of the test statistic is in the rejection region
- Reject H_0 if the p -value $\leq \alpha$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

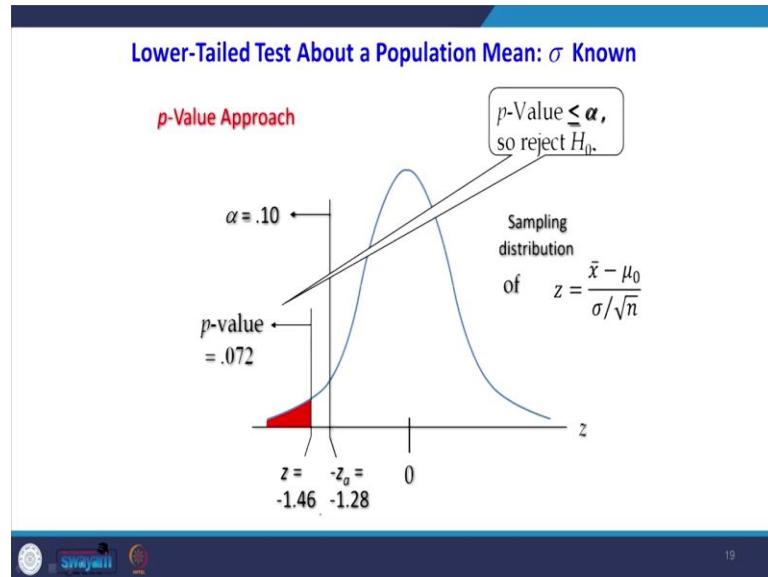
18

There are 3 approaches for hypothesis testing first approach is p-value approach most of the statistical package follow this method. Second method is critical value method, the third one is confidence interval value method. The confidence interval value method mostly used for 2-tailed test. First we will go for p-value approach that is a one-tailed hypothesis testing. What is the p-value the p-value is the probability computed using test statistic. You should be careful test the statistic that measures the support our lack of support provided by the sample for the null hypothesis.

So the p-value says whether it is supporting the null hypothesis or it is not supporting null hypothesis if the p-value is very high it will support null hypothesis you will accept null hypothesis. If the p-values be less it will not support null hypothesis we will reject the null hypothesis. Say we say that what is the test statistic the test statistic? For example in the Z context the test statistic nothing but this one $(\bar{X} - \mu) / (\sigma/\sqrt{n})$ that is the test statistic for Z test if it is a t-test this is $(\bar{X} - \mu) / (s/\sqrt{n})$.

So, $n - 1$ degrees of freedom, so, whatever value which have calculated with the help of sample that is called a test statistic, if the p-value is less than or equal to the value of the alpha then the value of the test statistics in the rejection region. I will show you in the next slide. Reject H_0 if the p-value is less than alpha. The p-value is very less it is not supporting null hypothesis here to reject it where alpha is significant level.

(Refer Slide Time: 21:36)



We will see how to use hypothesis, how to do hypothesis testing using p-value approach. The p-value approach the first one is, see assume that the problem alpha equal to 10% it is given this was alpha so we have to calculate this test statistic that that is your Z value. So, $X\bar{x}$ might be given $X\bar{x}$ is the sample mean, minus μ is the population mean what we have assumed Sigma value must be given, root of n.

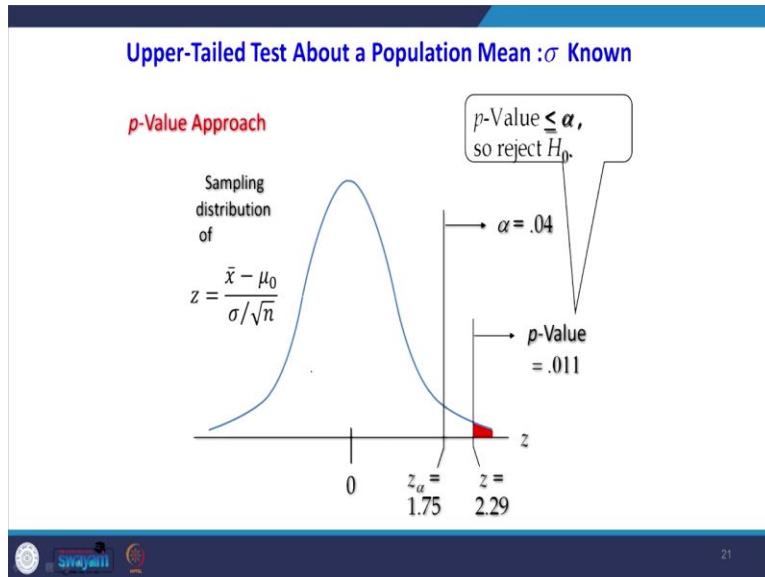
For example this value assumed that it is -1.46 okay. So, this -1.46 corresponding what is the left side area. So, this value is our p-value, okay how to get this one so when Z value is -1.46 we can get corresponding area of your normal distribution on the left hand side.

(Refer Slide Time: 22:39)

So, that one we can do with the help of Python for that first you have to import scipy so importing library from scipy import stats. Then the left side area is say, -1.46 the left side is Z statics -1.46 so when you put a minus this one stats.norm.cdf cumulative distribution function -1.46 we are getting the probability is 0.07 so that is nothing but 0.07 you see that this alpha is 10% so the p-value is less than the Alpha so way out to this region is a rejection region.

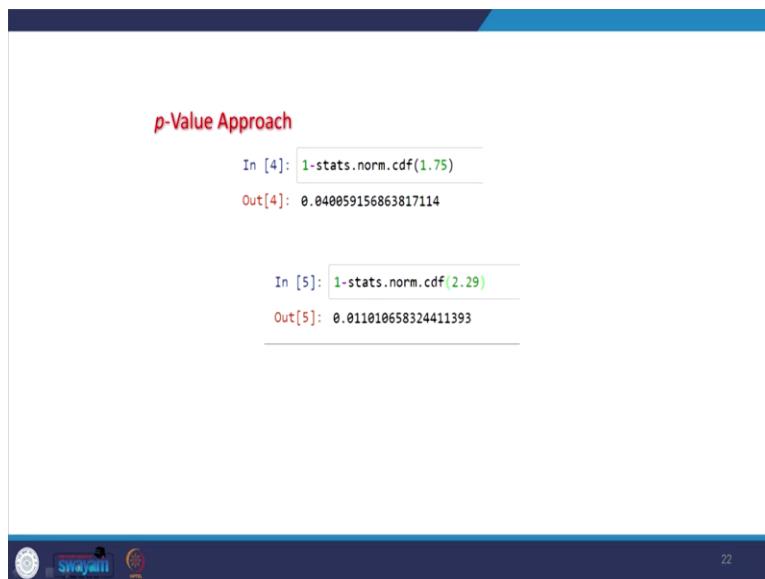
So, this region is acceptance region beyond this point it is the rejection region. So, the value of the P that is when Z value equal to -1.46 since we are standing on the left hand side that is we are standing on the rejection side we have to reject the null hypothesis. This is left side test lower tailed test this way explain.

(Refer Slide Time: 24:00)



Now suppose if it is a right tailed test say the calculated Z value is 2.29 we got some X bar value, μ value, Sigma by root n value, so suppose this is giving 2.29 so for alpha assume that alpha equal to 4% when alpha equal to 4, we go to mark it, alpha equal to 4% from the right to left. So, when alpha equal to 0.4% when Z values 2.29 we would what is the corresponding area towards the right side so what you have to do we can find out 1.75 also see that when Z values 2.29.

(Refer Slide Time: 24:46)



So `stat.norm.cdf` will give you the right side area when you put `1 - stats.norm.cdf(2.29)` will give you the left side area. So, this first one actually it is not required here because alpha will be directly given here this is only for proof this is for testing, how to use from Z, how this is Z value

from that we have find out the probability value. So, now the Z value is 2.29 so we want to know the right side area so 1 minus corresponding area that will give the right side area that 0.011.

So, this area I am saying this area is 0.011 now look at this alpha so the p-value is less than the alpha otherwise you see the p-value so this side is the rejection region this side the acceptance region. So, when the p-value is 0.01 still you are standing in the rejection region so how to reject a null hypothesis. In case if the p-value is 0.05 you might crossed the boundary after crossing the boundary you will be landing on the acceptance region so we have to accept a null hypothesis.

(Refer Slide Time: 26:07)

Critical Value Approach to One-Tailed Hypothesis Testing

- The test statistic z has a standard normal probability distribution.
- We can use the standard normal probability distribution table to find the z -value with an area of α in the lower (or upper) tail of the distribution.
- The value of the test statistic that established the boundary of the rejection region is called the critical value for the test.
- The rejection rule is:
 - Lower tail: Reject H_0 if $z \leq -z_\alpha$
 - Upper tail: Reject H_0 if $z \geq z_\alpha$

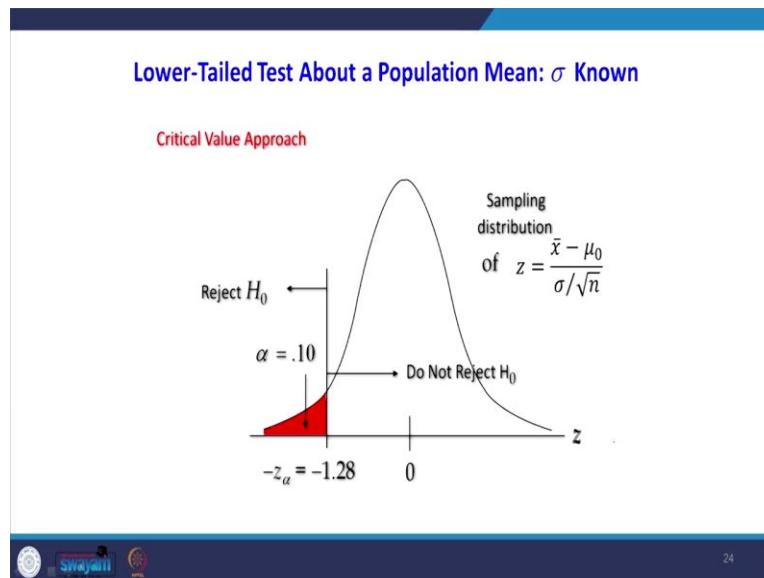
I will go to another method critical value approach for one tailed hypothesis are testing. The test statistic Z has the standard normal probability distribution we can use the standard normal probability distribution table to find out the Z value with an area of alpha in the lower tail or upper tail of the distribution. For example we know the Alpha value, say alpha value is 0.05, so this side area is 0.05 with the help of Python when alpha is 0.05 you can get the Z value this is lower tailed test.

For upper tail test when alpha equal to 0.05 you can find out corresponding Z value. In Python what you have to do if you want to know this right side, upper tail test you want to know the Z value you have to is $1 - 0.05$ for that probability you were to find out corresponding Z value. The

value of the test statistic that established the boundary of the rejection region is called critical value of the test. If it is for the 5% age you will get 1.645 here also -1.645.

So, this -1.645 is called a critical region. Rejection rule if it is a left tailed test reject if the Z value that means your calculated Z value is less than this -1.645 because you will be standing on the rejection side. If it is a right tailed test the calculated Z values greater than your table value then you have to reject it.

(Refer Slide Time: 27:57)

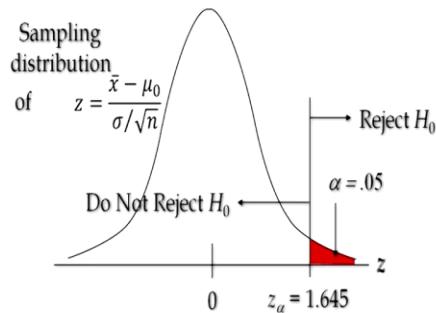


See for example Sigma is known the Alpha equal to 10% when alpha equal to 10% corresponding Z value is -1.28 this is our critical region this is our critical region. So, with the help of sample data you have to find out the Z value if the Z value is lying on this side you have to reject it. If the Z value is lying on that side you have to accept it.

(Refer Slide Time: 28:25)

Upper-Tailed Test About a Population Mean: σ Known

Critical Value Approach



26

For example when area equal to 0.1 the corresponding Z value is -1.28 and going back to previous slides -1.28. Now we will go for upper tail test when alpha equal to 0.05 so the right side area is 0.05 so this side is 0.95, so when the left side here is 0.95 corresponding Z values 1.645. If any calculated Z if this Z value is lying on this side for example 1.7 you have to reject the null hypothesis but is lying this side you have to accept the null hypothesis.

(Refer Slide Time: 29:01)

Steps of Hypothesis Testing

Critical Value Approach

- *Step 4. Use the level of significance α to determine the critical value and the rejection rule.
- *Step 5. Use the value of the test statistic and the rejection rule to determine whether to reject H_0 .



29

For example 0.95 the value is 1.65 what is the see how, now I now we have done hypothesis testing with help of p-value approach and critical value approach what is important point you to note that is the decision whether to reject or accept a null hypothesis in p-value method is decided by comparing the probability. Probability of what is the probability P value versus alpha

value. But in your critical value approach the decision is done by comparing the critical value and calculated Z value.

Decision will be same, only for comparison purpose sometimes we compare probability sometimes we compare critical value but the end result will be same. So, what is the first step will develop null and alternative hypothesis step 2 specify the level of significance alpha this is very important. Before starting of the test you have to decide the significance level. Step 3 collect the sample data and compute the test statistic.

This test statistic maybe your t value or it may be z value. The p-value approach what will you do use the value of test statistics to compute the p-value, if the p-value is less than or equal to alpha you rejected the same step for critical value method use the level of significance alpha, to determine the critical value and rejection rule. Use the value of test statistic and the rejection rule to determine whether to reject H_0 .

Dear students in this lecture so far what we have seen we have seen what is hypothesis what does the null and alternative hypothesis. We have learnt how to formulate hypotheses then we have seen hypothesis testing. In the hypothesis testing we have seen what is left tail test, what is the right tail test. What is the two tail test then we have seen the theory of how to test the hypothesis by using p-value approach and by using the critical value approach.

The next class will take one problem with the help of that problem will formulate the hypothesis then we will test the hypothesis with help of p-value approach and critical value approach then we will compare the result. And one more thing one more method that I did not cover in this lecture is that is testing the hypothesis with the help of confidence interval that we will do in the next class, thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 17
Hypothesis Testing- II

Welcome students in the last class we have seen how to formulate the hypothesis then we have seen some theory when the hypothesis should be accepted when the hypothesis should be rejected. In this class we will take some practical examples then we will solve them then we will understand the concept of hypothesis in detail. So, the class objective is when the population standard deviation is known how to do the hypothesis testing.

(Refer Slide Time: 01:00)

One-Tailed Tests About a Population Mean: σ Known

- Example: The mean response times for a random sample of 30 Pizza Deliveries is 32 minutes
- The population standard deviation is believed to be 10 minutes.
- The pizza delivery services director wants to perform a hypothesis test, with $\alpha = 0.05$ level of significance, to determine whether the service goal of 30 minutes or less is being achieved.



SWAYAM

Here hypothesis testing is we are going to check the population mean will take own problem this is example of one tailed test about the population mean when Sigma is known Sigma means population standard deviation. The assumption is on population mean so the problem is the mean response times for the random sample of 30 pizza deliveries is 32 minutes. So, they conducted a sample survey in that they have sample sizes 30 out of 30 they found that the mean delivery time for Pizza is 32 minutes.

The population standard deviation is believed to be 10 minutes Sigma is known this population standard deviation this population standard deviation is nothing but your Sigma 10 minutes this

is nothing but over Sigma. The pizza delivery services director wants to perform a hypothesis test when alpha equal to 0.05 level of significance to determine whether the service goal of 30 minutes or less is being achieved.

Manager of that store or that shop wanted to verify whether the survey's goal of 30 minutes or less is being achieved or not. So, now the status quo first thing is formulating the hypothesis the status quo is the Pizza is delivered within 30 minutes.

(Refer Slide Time: 02:19)

Given Values

<ul style="list-style-type: none">• Sample• Sample mean = 32 Min• Sample size = 30	<ul style="list-style-type: none">• Population• $\alpha = 0.05$• Population mean = 30 Min
--	--

Before that we will see what are the values are given there are any hypothesis testing there is a two kind of data some data from sample some data from population. So, in the sample, sample mean is 32 minutes sample sizes n. so, this sample mean this is nothing but your \bar{x} this is nothing but your n, so with respect to population, population mean which we have to assume is the 30 minutes what is the population standard deviation also has to be given, what is the population standard deviation? Sigma equal to going back yeah the population standard deviation is 10 minutes.

(Refer Slide Time: 03:05)

One-Tailed Tests About a Population Mean:

σ Known

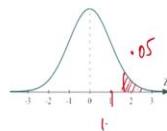
1. Develop the hypotheses.
2. Specify the level of significance.
3. Compute the value of the test statistic.

$$H_0: \mu \leq 30$$

$$H_a: \mu > 30$$

$$\alpha = .05$$

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{32 - 30}{10 / \sqrt{30}} = 1.09$$



6

Now first we will solve this problem using p-value approach what is the step 1 develop the hypothesis develop the hypothesis in the previous class also I given some hint the status quo, status quo should go to null hypothesis. What is the status quo currently the pizza is delivered and the average of 30 minutes, so mu is less than or equal to 30. After you write the null hypothesis then you should go for alternative hypothesis the clue used that these signs are complementary when you write for null hypothesis it is a less than or equal to for alternative episodes it should be greater so greater than 30.

The step 2 is specified the level of significance alpha is given 5% now we have to decide whether it is a right tailed test or left tailed test. As I told you by looking at the sign off your alternate hypothesis it is greater than 30 so it is a right tailed test for example this is right tailed test. What is alpha it is 0.05. The next one compute the value of test statistic for the test statistic $Z = (\bar{X} - \mu) / (\sigma / \sqrt{n})$, \bar{X} bar is given 30 to μ (mu) is as you would mean divided by 10 it is the population standard deviation/ root of n, n is your sample size, $z = 1.209$. So when you mark 1.09 approximately it will be here 1.09.

(Refer Slide Time: 04:40)

One-Tailed Tests About a Population Mean: σ Known

p-Value Approach

4. Compute the p-value,

For $z = 1.09$, $p\text{-value} = 0.137$

5. Determine whether to reject H_0 .

- Because $p\text{-value} = 0.137 > \alpha = 0.05$, we do not reject H_0 .
- There are not sufficient statistical evidence to infer that Pizza delivery services is not meeting the response goal of 30 minutes.



8

So, what we have to do then the calculated Z value are the test statistics is 1.09 we have to find out what is the right side area that is whatever p-value, so what is the meaning is when it is Z values 1.09 so we are to suppose this is 1.09 we have to mark this side area that side area is nothing but p value p value. So, with the help of Python when you go for 1 minus stat.norm.cdf 1.09 because cdf is we are finding minus infinity to Z value when you want the right side area that has to be divided by 1 that is nothing but if I draw here one more time in Python we can get area when Z values 1.09.

For example approximately here we have to find out what is the right side the area. So, for finding the right side area if you put stat.norm.cdf of 1.09 because the Python is giving area from minus infinity to here Z value. So, you will get the left side area but we want to know the rights area. So, since you know the area is 1, so 1 minus this left side area will give you the right side area. So, the right side area this one is 0.137 you have to compute the p-value for Z equal to 1.09 the p-value is 0.137.

Now we have to determine whether to reject H_0 or not what has happened since the p-value 0.137 is greater than alpha value because alpha is 0.05 so the p-value is greater than alpha value so we do not reject null hypothesis. So, what is the meaning is, again I am drawing one more time even though there are many places and drawing normal distribution that will be for our

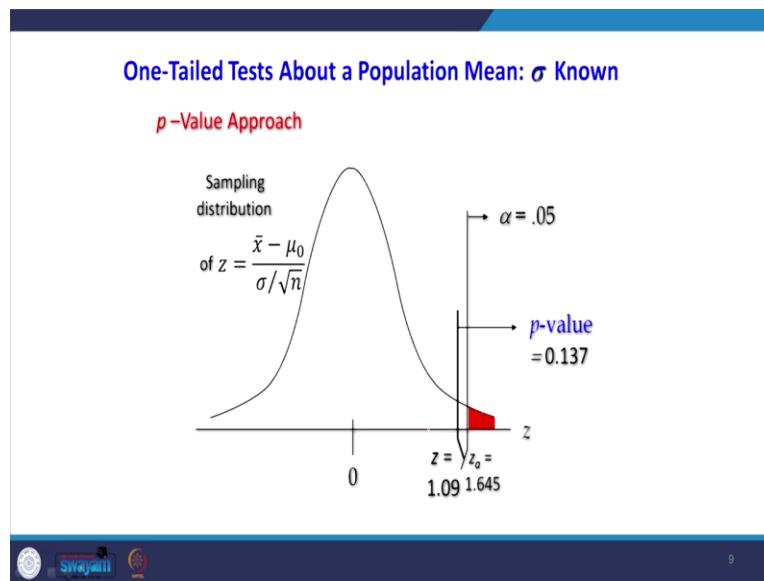
understanding purpose, this is 0.05. So, whatever value which is on the right side of this 0.05 will be rejected what happened the area which I found this area is 0.137.

So, this area is 0.137 so now what happened now I have entered into the acceptance region. So, I have to accept the null hypothesis. In case with the p-value is 0.04 so I will be standing here because this is up to this I am writing this is 0.05 when you say 0.04 I will be standing here so that means I am standing in the rejection region I have to reject it. Now what has happened that we are crossing that boundary of 0.05 so we have entered into the acceptance region.

So we have to accept the null hypothesis but generally we would not say accept do not reject null hypothesis. So, what is the conclusion there are not sufficient statistical evidence to infer that Pizza delivery service is not meeting the response goal of 30 minutes. So, when you say accept null hypothesis so what we say that the μ is less than or equal to 30 minutes that means the Pizza is delivered before 30 minutes.

You know that offer is there if it is not delivered within 30 minutes the Pizza is free for you. So, they make sure that all deliveries are delivered within 30 minutes.

(Refer Slide Time: 08:12)



The same example you see when Z equal to 1.09 corresponding p value is 0.137 but the Alpha is 0.05 the red region represents the rejection region. So, what has happened we cross into the acceptance region so we have to accept the null hypothesis.

(Refer Slide Time: 08:32)

One-Tailed Tests About a Population Mean: σ Known

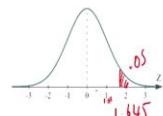
Critical Value Approach

4. Determine the critical value and rejection rule.

- For $\alpha = .05$, $z_{.05} = 1.645$
- Reject H_0 if $z \geq 1.645$

5. Determine whether to reject H_0 .

- Because $1.645 \geq 1.05$, we do not reject H_0 .



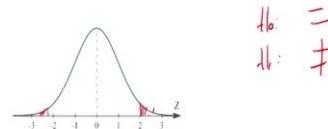
11

Then the same problem will do with the help of critical value approach in both approach we have to get the, we have to get the same answer. So, we will continue from the step 4 first determine the critical value and rejection rule. So, what is the rejection rule is when alpha equal to 0.05 we have to find out what is the Z value Z value is 1.645. Then you see that our calculated Z value is one 0.05 that is nothing but our test statistic. So, test statics will statistic will be 1.05 here 1.05.

So, what is the logic if the calculated Z value are the test statistics is, if it is falling on the rejection reason we have to reject it but here it is falling on the acceptance region so we have to accept the null hypothesis.

(Refer Slide Time: 09:31)

p-Value Approach to Two-Tailed Hypothesis Testing



12

The previous example was the p value for one tailed test now we will see how to do hypothesis testing for a two-tailed test as I told you how to know it is the two-tailed test in alternative hypothesis if the sign is for example H_0 is this one H_0 if the sign is not equal to then it is a two-tailed test. Generally two tailed test what will happen there will be an upper limit there will be a lower limit.

If any values any test statistics if it is false on the above this upper limit of the acceptance region we will reject it for this falls below the below the acceptance region we will reject it.

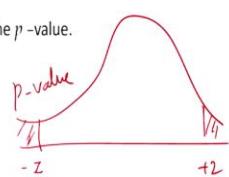
(Refer Slide Time: 10:16)

Compute the p-value using the following three steps:

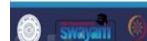
1. Compute the value of the test statistic z .
2. If z is in the upper tail ($z > 0$), find the area under the standard normal curve to the right of z .
3. If z is in the lower tail ($z < 0$), find the area under the standard normal curve to the left of z .
4. Double the tail area obtained in step 2 to obtain the p -value.

The rejection rule:

Reject H_0 if the p-value $\leq \alpha$.



13



Now computing the p-value using the following 3 steps one is compute the value of test statistic Z if Z is an upper tail find the area under the standard normal curve to the right of the Z. If the Z is a lower tail that is if the less is less than 0 find the area under the standard normal curve to the left of Z. So, double the tail area obtained by in step two that is a logical why what we are going to do, since it is a two tailed test.

So, whatever area which were found left side or right side that has to be doubled to obtain the p-value. The rejection rule is if the double devalue double the p-value is less than or equal to alpha reject it otherwise accept it so what is the what it say is that you go this way for a test statistics, for the test statistics for example Z you find what is the area you multiply this left side area this is a p-value multiplied by 2 times because it is a two-tailed test.

And not only that it is a symmetric if after multiplying if the p-value is still less than or equal to alpha we have to reject it. Otherwise what you can do instead of multiplying you when it is a minus Z you find out what is the p-value when it is the plus Z what is the find of p-value you add it the added p-value should be less than or equal to; if it is less than or equal to alpha we have to reject it.

(Refer Slide Time: 11:53)

Critical Value Approach to Two-Tailed Hypothesis Testing

- The critical values will occur in both the lower and upper tails of the standard normal curve. $\alpha/2$
- Use the standard normal probability distribution table to find $z_{\alpha/2}$ (the z-value with an area of $\alpha/2$ in the upper tail of the distribution). $\alpha/2$
- The rejection rule is:

Reject H_0 if $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$.

14

The critical value will occur in both lower and upper tail of the standard normal curve use the standard normal probability distribution table to find out $Z_{\alpha/2}$. Why we are doing $Z_{\alpha/2}$ because a

2 tail if alpha equal to 5% for example so $\alpha/2$ is 2.5% it is 0.025 so when α by 2 is 0.025 we have to find out corresponding Z value on left side and right side. So, the rejection rule is if Z is less than the lower limit reject it or if Z is above the upper limit reject it.

(Refer Slide Time: 12:45)

Two-Tailed Tests About a Population Mean: σ Known

- Example: Milk Carton
- Assume that a sample of 30 milk carton provides a sample mean of 505 ml.
- The population standard deviation is believed to be 10 ml.
- Perform a hypothesis test, at the 0.03 level of significance, population mean 500 ml and to help determine whether the filling process should continue operating or be stopped and corrected.

An illustration showing three blue and white milk cartons standing upright. The carton on the right has a purple circle with a question mark inside it drawn on its side.

 Swayam 

15

Here the Z means sample statistic we will do an example for that his example is for doing hypothesis testing for the two-tailed test when Sigma is known. The example is a milk carton assume that a sample of 30 milk carton provides a sample mean of 505 ml, the population standard deviation is believed to be 10 ml, perform a hypothesis test are at 0.03 level of significance when the population mean is 500 ml. To help to determine whether the filling process should be continued to operating or it has to be stopped and corrected.

So, what is happening there is a assume that it is assembly line so the in the assembly line that the bottles are filled with 500 ml, what is happening generally if it is over filling also there is a problem if it is under filling also there is a problem that is why if it is $H_0: \mu = 500$ ml, $H_a: \mu \neq 500$ ml. The logic why we did not go for left tail or right tail test is because we have to go for not equal to because even over filling and under filling is the problem for us that is why we should go for two-tailed test.

(Refer Slide Time: 14:08)

Two-Tailed Tests About a Population Mean:

σ Known

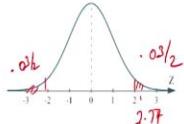
1. Determine the hypotheses.
2. Specify the level of significance.
3. Compute the value of the test statistic.

$$H_0: \mu = 500$$

$$H_a: \mu \neq 500$$

$$\alpha = .03$$

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{505 - 500}{10 / \sqrt{30}} = 2.74$$



18

So, what are the data given here as usual data will be given for sample and population n equal to 30 this sample mean is 505 ml with respect to population what kind of data is given we are assuming that mu equal to 500 and standard deviation Sigma equal to 10 ml and alpha equal to 0.03. This problem that is a two tailed problem will solve with the help of p-value approach. First you have to determine the hypothesis.

You see that mu equal to 500 why this as I told you because both overfilling and under filling will cause the problem for the company. So, the hypothesis is formulated specify the level of significance alpha it is given in the problem it is 0.03% it is a two-tailed test, so I have to mark this side 0.03 by 2 left side also it is 0.03 by 2. Next I have to compute Z statistic, Z statistic is $(\bar{X} - \mu) / (\sigma / \sqrt{n})$, \bar{X} bar is 505 mu is assumed mean 500 divided by 10 is given root of 30, so 2.74 so you have to mark this 2.74.

(Refer Slide Time: 15:50)

```
In [9]: 1-stats.norm.cdf(2.74)
Out[9]: 0.003071959218650444
```



```
In [10]: (1-stats.norm.cdf(2.74))*2
Out[10]: 0.006143918437300888
```

For example assume that the 2.74 is here so what I have to do when the Z value is 2.74 we have to find out the area towards the right in Python if you type this $1 - \text{stat.norm.cdf}$ of 2.74 the right side area is 0.003. If you multiply this both side because multiply two times because it is symmetric. So, this what this meaning is in Python it is this way so when Z equal to 2.74 corresponding right side the area is point 0.00307.

This side also when Z equal to minus 2.74 the area is 0.00307 you mu add both you will get 0.0061.

(Refer Slide Time: 16:42)

Two-Tailed Tests About a Population Mean: σ Known

p –Value Approach

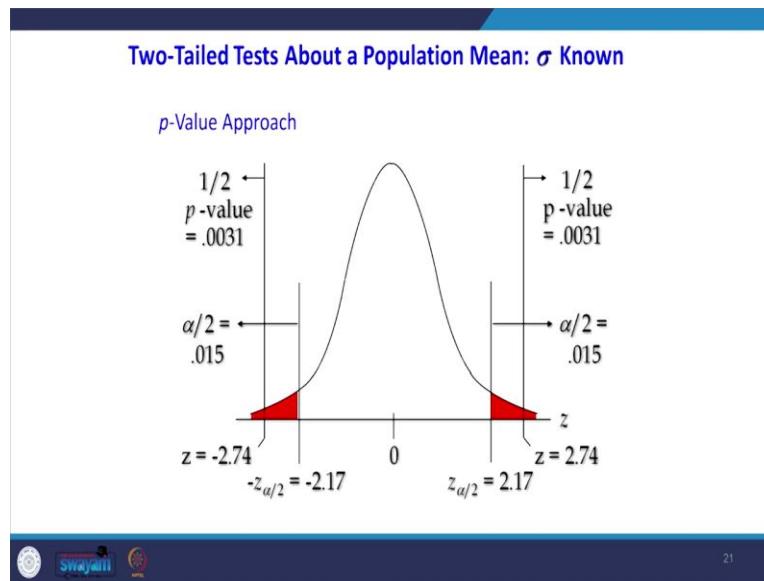
4. Compute the p –value.
 - For $z = 2.74$, $p\text{-value} = 2(1 - .9969) = .0061$
5. Determine whether to reject H_0 .
 - Because $p\text{-value} = .0062 < \alpha = .03$, we reject H_0 .

There is no sufficient statistical evidence to infer that the null hypothesis is true (i.e. the mean filling quantity is not 500 ml)

So what is happening that 0.0061 is less than your 0.03 see this 0.0061 this was 0.0062 by after approximation the Alpha is 0.0 there is still it is that less than the alpha value so we have to reject H₀. When we reject H₀ there is no sufficient statistical evidence to infer that the alternative hypothesis that means the mean filling quantity we are rejecting null hypothesis so when we reject null hypothesis what was our null hypothesis $\mu = 500$, H₁: $\mu \neq 500$, when you reject it there is no sufficient statistical evidence to infer that the alternative hypothesis is true.

So we have found that the p-value is 0.002 that is less than alpha 0.03, so we have to reject null hypothesis, when you reject a null hypothesis we are accepting our alternative hypothesis that there is no sufficient statistical evidence to infer that the null hypothesis is true. So, the mean filling quantity is not 500ml so immediately they have to stop the assembly line they had to make the corrective action that is the inference.

(Refer Slide Time: 17:57)



Yes that was shown here in the picture form Z equal to 2.74 is the test statistics. So, the right side area is 0.0031 when test two statistics is -2.74 the left's idea is 0.0031 after adding that still it is less than or equal to alpha we have to reject it otherwise we can compare this 0.0031 versus 0.15 the p-value this is half of the significant value, the half of the p-value is 0.03 that is lesser than the 0.015 so we can reject it. But many software packages you may not give this half of the p-value and half of the Alpha value.

You will get the added value that means this 0.0031 is added with another 0.0031 then this alpha on the 0.015 is added with another 0.015 so the added p-value is compared with added alpha value then we take the decision if the p-value is less than alpha we are to reject it.

(Refer Slide Time: 19:08)

Two-Tailed Tests About a Population Mean : σ Known

- **Critical Value Approach**

4. Determine the critical value and rejection rule, for $\alpha/2 = .03/2 = .015$, $z_{.015} = 2.17$

Reject H_0 if $z < -2.17$ or $z > 2.17$

5. Determine whether to reject H_0 .

Because $2.74 > 2.17$, we reject H_0 .

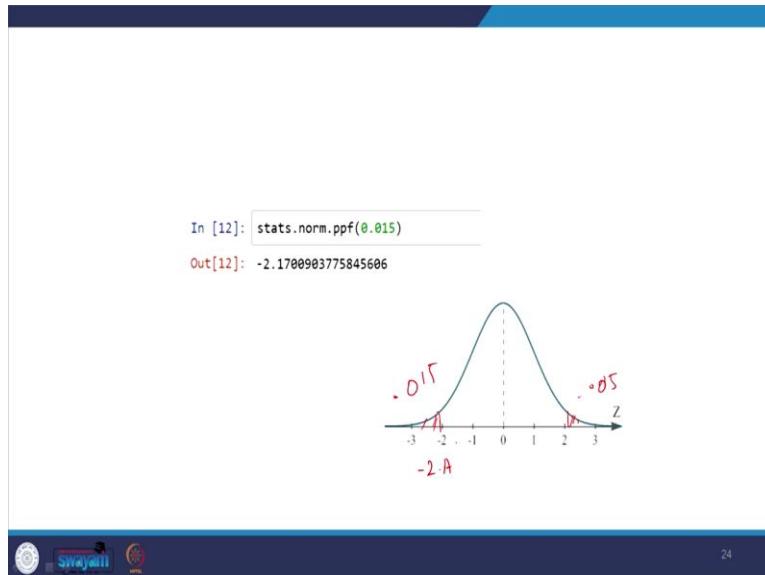
There is sufficient statistical evidence to infer that the null hypothesis is not true

So, here we are rejecting I will go for critical value approach the critical value approach will continue from the fourth step determine the critical value and the rejection rule for alpha by 2 0.015 so what is the meaning is when alpha is 0.015 we have to find out this critical value for the right side when this side area is 0.015 we were to find out - Z critical value. So, if the calculated Z value is lying on right side we are projected to what is lying on the left side we have to reject.

Because what happened the 2.74 is our calculated Z value otherwise test statistic sample statistic this value is 2.17, so the 2.74 will be on this side 2.74 will be on the rejection side. So, we have to reject it so there is a sufficient statistical evidence to infer that the alternative hypothesis is not true. Now test statistics 2.74 lying on the rejection side we have to reject our H_0 , so the conclusion is there is sufficient statistical evidence to infer that the null hypothesis not true.

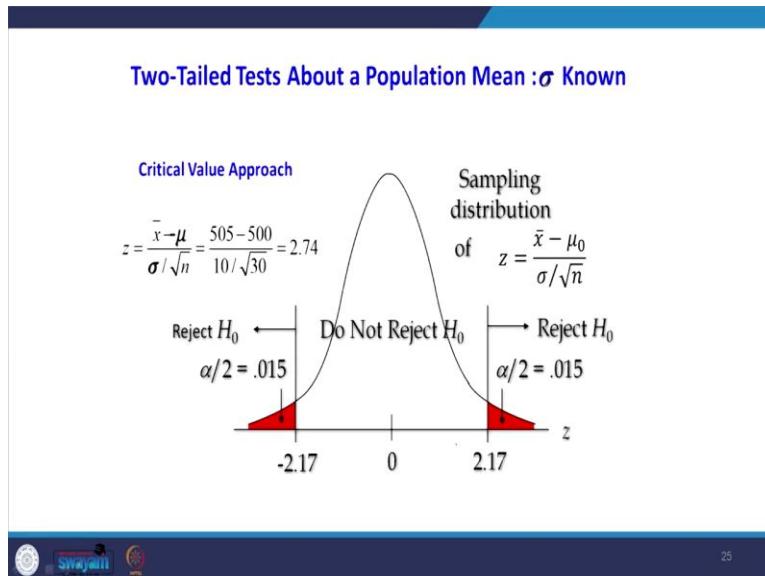
So we have to accept our alternative hypothesis that means the assembly process is not filling average value of 500 ml. So, we have to stop that assembly line then we have to make corrective actions.

(Refer Slide Time: 20:50)



So, what is the step here as I told you when this value is alpha by 2 that is a 0.015, so the corresponding, this is a positive side the left side 0.015 so the left side -2.17 how can you get this one when you type stats.norm.cdf of in Python when you put 0.015 you will get lower limit of our critical value this is symmetric. So, right side, it is also will be same value.

(Refer Slide Time: 21:26)



This also same thing what has happened when alpha by 2 is 0.015 the lower limit is -2.17 it is 0.15 on the right hand side the upper limit is 2.17 this Z value that is we calculated 2.74 so this 2.74 will be this side 2.74 will be on the rejection side you have to reject it. In case for example the Z value, say 2 for example 1.5 say 1.5 will be here so we have to accept it.

(Refer Slide Time: 22:02)

Confidence Interval Approach to Two-Tailed Tests About a Population Mean

The 97% confidence interval for 500 is

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 505 \pm 2.17 \frac{10}{\sqrt{30}} = 505 \pm 3.9619 \\ = 501.03814, 508.96186$$

Because the hypothesized value for the population mean, $\mu_0 = 500$ ml, is not in this interval, the hypothesis-testing conclusion is that the null hypothesis, $H_0: \mu = 500$, is rejected.



28

We will solve the same problem with the help of confidence interval approach confidence interval approach for two tailed test about the population mean. So, select the simple random sample from the population and use the value of the sample mean X bar here X bar to develop your confidence interval for the population mean μ (μ), if the confidence interval contains hypothesis value of 500 do not reject it.

So how we are going to develop this conference interval is we know this very familiar formula $(x\bar{} - \mu)/(\sigma/\sqrt{n})$ when you readjust this so we can find out in terms of $x\bar{}$ the upper limit of μ and lower limit of so is, $\mu + Z(\sigma/\sqrt{n})$ will be the upper limit when you put $\mu - Z(\sigma/\sqrt{n})$ will be the lower limit. So, what we are do with the help of $x\bar{}$ we have to express the upper limit lower limit of μ so that formula has come from this $Z = (X\bar{} - \mu)/(\sigma/\sqrt{n})$.

So the upper limit of μ is $X\bar{} + Z \Sigma \sqrt{n}$ lower limit will be $X\bar{} - Z \Sigma \sqrt{n}$ this value by adjusting this Z equations we got this one in this interval. Suppose we have to find out the upper limit say this is lower limit this is upper limit, in this interval if the 500 is the assumed mean is lying we have to accept null hypothesis. Otherwise reject it, actually H_0 should be rejected if μ happens to be equal to one of the endpoint of the conference interval.

Now you see that the formula which have explained previous slide $x\bar{} \pm Z_{\alpha/2} (\sigma/\sqrt{n})$, and $X\bar{}$ bar is $505 \pm Z_{\alpha/2}$ is 2.17, so Sigma and n so sample size is 30 so this is 505 ± 3.9619 so the lower

limit is this is lower limit this is upper limit. So, in this interval we are not able to capture 500, so we have to reject null hypothesis, so to see this because the hypothesis value for the population mean μ 500ml is not in this interval so not in this interval the hypothesis testing conclusion is that the null hypothesis $H_0: \mu = 500$ is rejected.

Dear students what you have seen in this lecture so far we have taken one practical problem for the pizza delivery problem. We have learned how to test one tailed test that is left tail test then we have learnt how to do the two tail test. In one tail test first we solved with help of p-value approach then we solved with the help of critical value approach. In two tail test also first you solved with the help of p-value then critical value.

Then the third one which ever solved using confidence interval method in all these 3 method the final result is same that we have rejected our null hypothesis. In the next class will start with the t-test so what will happen in t-test so far we the population standard deviation is given there may be a situation where you may not know the population standard deviation that time you should go for t-test that will continue in the next class.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 18
Hypothesis Testing- III

Welcome students in the previous lecture we have seen hypothesis testing when Sigma is known so that test we call it as Z test. Now we will go to be another category of hypothesis testing procedures where Sigma is not known. Most of the time the population standard deviation is not known to us so that time we should go for another test that is called t test.

(Refer Slide Time: 00:49)

Tests About a Population Mean: σ Unknown

- Test Statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

This test statistic has a t distribution with $n - 1$ degrees of freedom.

SWAYAM

2

I will explain what is the connection between the Z test and t test? So, previously you remember that we have used the $Z = (\bar{X} - \mu) / (\sigma/\sqrt{n})$. So, in this case since Sigma is not known to us instead of Sigma that Sigma is replaced by s that is a sample standard deviation. The another relation between Z and t is when the n is increasing look at this picture here the blue one is Z distribution the pink one is t distribution.

So, when n is increasing when the sample size is increasing the behavior of Z test and t test is same that is why in many software packages there would not be separate tab for doing Z test there will be a tab only for doing t test for example hypothesis testing and in SPSS. When you go

for even in Minitab also when you go for that there would not be a column to do the Z test but there will be a column for t-test.

For doing Z test the t test is enough, so what is the t, $t = (\bar{X} - \mu) / (s / \sqrt{n})$. here the degrees of freedom will come into place because the shape of the t distribution it is affected by the degrees of freedom when the degrees of freedom is increasing so the behavior is same. So, the test statistic has a t distribution with $n - 1$ degrees of freedom.

(Refer Slide Time: 02:22)

Tests About a Population Mean: σ Unknown

Rejection Rule: p -Value Approach

Reject H_0 if $p\text{-value} \leq \alpha$

Rejection Rule: Critical Value Approach

$H_0: \mu \geq \mu_0$ Reject H_0 if $t \leq -t_{\alpha}$

$H_0: \mu \leq \mu_0$ Reject H_0 if $t \geq t_{\alpha}$

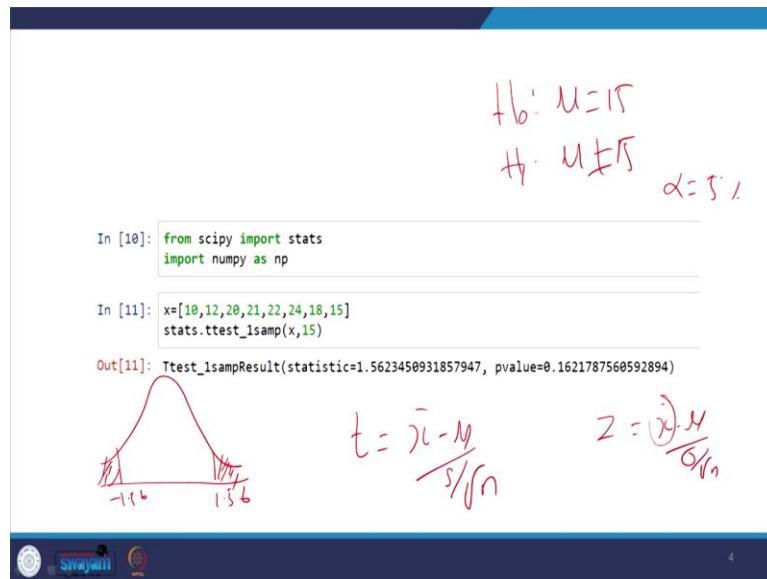
$H_0: \mu = \mu_0$ Reject H_0 if $t \leq -t_{\alpha/2}$ or $t \geq t_{\alpha/2}$

What is the rejection rule the same whatever we have seen for the previous lecture that is Z test that rule is applicable for here also so reject H_0 if the $p\text{-value} \leq \alpha$, if it is a critical value approach suppose if $\mu \geq \mu_0$ so alternative hypothesis will $\mu < \mu_0$ so it is a left tailed test. So, in the left tailed test, test statistic t is less than your $-t_{\alpha}$, you have to reject it.

For example see here this is $\mu \leq \mu_0$ so the signs are complementary so this right tailed test so what is this left tailed test. So, the left tailed test will be like this so this is t_{α} , so if the calculated t value is lying on this side we are to reject it what does this right tail test I am writing this side so if it is this way this is right if the t is like this is t_{α} the t , calculated t is lying on the right side we have to reject it if $\mu = \mu_0$ what will be the alternate hypothesis here $H_1: \mu \neq \mu_0$.

So, this will be this way there will be $t_{\alpha/2}$ on the right side $t_{-\alpha/2}$ on the left side if the t value calculated t value is lying on any of this side we have to reject our null hypothesis.

(Refer Slide Time: 03:55)



For doing t test in Python we have to import stats, from scipy import stats then you have to import numpy else import numpy as np, now X equal to 10 12 20 21 22 24 18 and 15 the function for doing t test is start stat t test underscore one sample here the X represent this array the mu represents our as you would mean. So, for this problem here the null hypothesis is $H_0: \mu = 15$, $H_1: \mu \neq 15$.

So when you run this you are getting this test statistics what is this value does $t = (X \bar{ } - \mu)$ divided by (s / \sqrt{n}) . So, what has happened the Python calculated the value of $X \bar{ }$ from this given array and the value of s sample standard deviation with the help of $X \mu$ this is 15 n is the sample size it is taken care. So, what is happening the previously what you are then when you are doing Z test we are using $(X \bar{ } - \mu)$ divided by (σ / \sqrt{n}) the $X \bar{ }$ has to be formed with from the sample.

But here you need not do that one just you mention that array name the built-in function will take care. So, this p value what we are getting is the two-sided p-value, two-sided p-value me in the sense suppose if it is a two-tailed test, so this site when t value what is the t value when t is 1.56 so plus 1.56, - 1.56, so the total area that is the right side area and this left side area that the area

is 0.16 assume that my alpha equal to say 5% what is happening the p value is exceeding the 5%. So we have to accept our null hypothesis this is the way to do the t test in Python.

(Refer Slide Time: 06:08)

One-Tailed Test About a Population Mean: σ Unknown
Example: Ice Cream Demand

- In a ice cream parlor at IIT Roorkee, the following data represent the number of ice-creams sold in 20 days
- Test hypothesis $H_0: \mu \leq 10$
- Use $\alpha = .05$ to test the hypothesis.



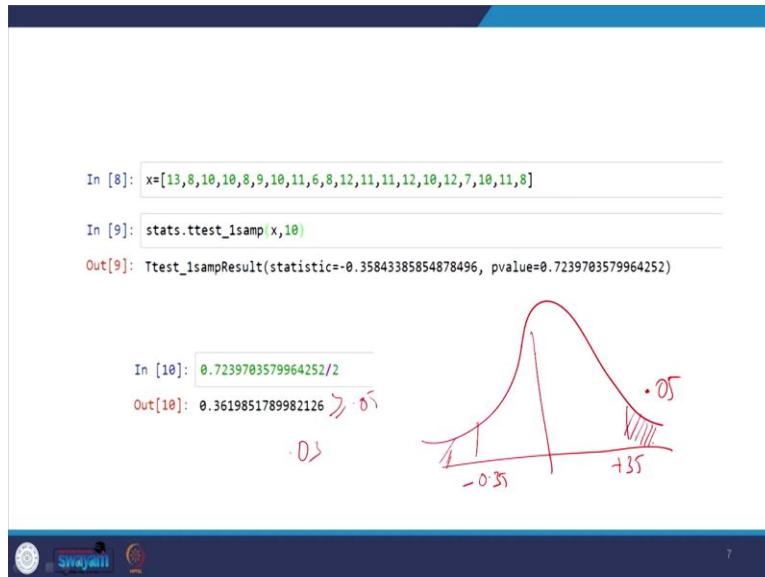
Day	No. of Ice-cream Sold	Day	No. of Ice-cream Sold
1	13	11	12
2	8	12	11
3	10	13	11
4	10	14	12
5	8	15	10
6	9	16	12
7	10	17	7
8	11	18	10
9	6	19	11
10	8	20	8

5

It will take an example for this in an ice cream parlor at IIT Roorkee the following data represents the number of ice cream sold in 20 days. So, here n equal to 20, the 20 days data were surveyed the shop owner want to test that the average is less than 10 by taking alpha equal to 5% so what is happening there are 20 data set is there the number of ice cream sold on day 1 is 13 day 2 is 8 and so on so for day 20, 80.

So, if you are doing manually that means with the help of statistical table we have to find out the sample you have to find out the $X_{\bar{}}$ bar then we have to find out the sample standard deviation then you have to use this formula $(X_{\bar{}} - \mu) / (s/\sqrt{n})$.

(Refer Slide Time: 06:58)

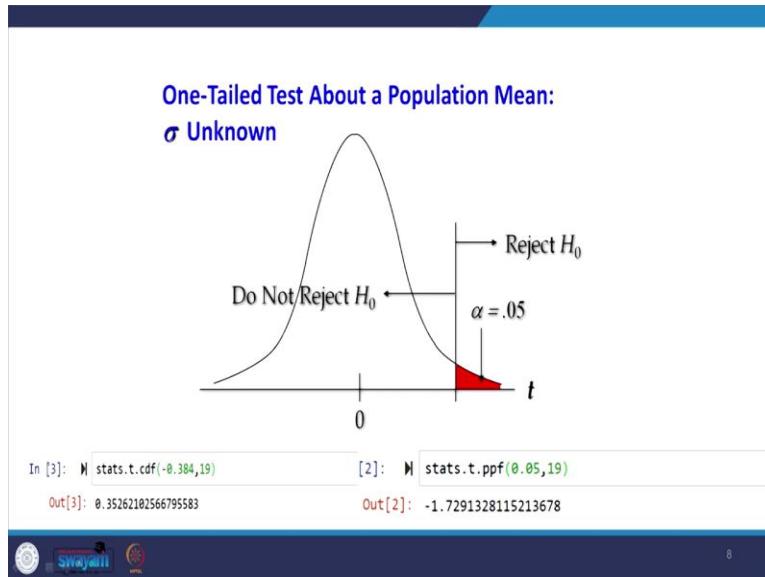


But in Python it is helping us very easily so we will go back, so what is happening the H_0 : is mu less than or equal to 10 so H_0 : mu less than or equal to 10. So, what will be over alternative hypothesis: mu greater than 10, so it is a right tailed test. So, right tailed test I have to shade to this side okay what is alpha is 0.05 first you will saw 0.05 so I have stored this given value into an object X so X is 13 8 10 all the values which I was shown in the previous table that I have stored in this one. Then stats dot t-test underscore 1 samp in bracket X ,10 the here the 10 is our assumed mean assumed population mean.

This X is this array, so what we are getting we are getting the t value is - 0.35 so I am drawing this distribution now, so what kind of yeah this is right tailed test right, right tailed it is 0.05 but when we do the sample statistics we are getting that is - 0.35 but this 0.72 is when Z equal to - 0.35 what is the left side area when Z equal to plus 35 what is the right side area? So, that added area is this 0.72, 0.723. Now since it is a one tailed test we have to divide by 2 when you divide this one it is 0.36 so this 0.36 is greater than 0.05.

So, we have to accept null hypothesis. In case if the p-value is for example 0.03 so we will be stand, will be landing on here so we have to reject the null hypothesis yeah this is right tailed test this is the right tailed test.

(Refer Slide Time: 09:20)



What was the t value here what was the t value one tailed test about the population mean when Sigma is unknown the previously the value of t is - 0.384, so when it is - 0.384 the corresponding p-value suppose if it is - 3.84 it will be here this will be - 0.384, - 0.384. so, this left side area is 0.3526 because in Python you see that from the t value you can find out the area stats dot t dot cdf, you have to find out the way to enter the t value and corresponding degrees of freedom.

Previously our sample size is 20 so the degrees of freedom is 19 so when the, this is calculated t value the corresponding area is 0.35. so, what is happening this 0.35 is greater than this is left side left side the area similarly if you put plus 0.384 you will get the area towards left that has to be substrate from – 1, so we will get the right side area. So, the right shady area will be 0.35 approximately 0.35 will be right side area will be here 0.35. So 0.05 is less so the p-value is more than the Alpha so we have to accept the null hypothesis.

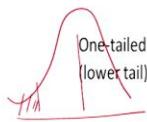
(Refer Slide Time: 10:57)

Null and Alternative Hypotheses: Population Proportion

- The equality part of the hypotheses always appears in the null hypothesis.
- In general, a hypothesis test about the value of a population proportion p must take one of the following three forms (where p_0 is the hypothesized value of the population proportion).

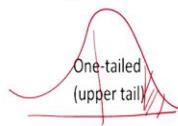
$$H_0: p \geq p_0$$

$$H_a: p < p_0$$



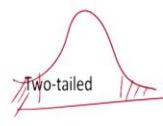
$$H_0: p \leq p_0$$

$$H_a: p > p_0$$



$$H_0: p = p_0$$

$$H_a: p \neq p_0$$



10

Next we will go for hypothesis testing for proportion similar to null and alternate hypothesis for mean. So, here the Equality part of the hypothesis always appear in the null hypothesis in general a hypothesis test about the value of the population proportion P this is P population proportion must take one of the following 3 forms right. So, for example the $P \geq P_0$ so this is this is a situation like this; what is happening this is this is a left tailed test example for this.

There is another possibility it may appear this way this is your right tailed test. How I am naming left tail or right tail, I'll write a test by looking at the sign of your alternate hypothesis. Now it is a two tailed test what would be two tailed test? This way if anything below all, so will reject it anything above it is, similar to what you have seen previously.

(Refer Slide Time: 12:03)

Tests About a Population Proportion

Test Statistic

$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}}$$

where: $\sigma_{\bar{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$

assuming $np \geq 5$ and $n(1-p) \geq 5$



11

Here the test statistic test statistic raised $Z = (\bar{p} - p_0)$ divided by $\sigma_{\bar{p}}$, here the \bar{p} is your sample proportion this is your assumed a population proportion this is a standard error for the proportion. So, the standard error is $\sqrt{(pq)/n}$ that is $\sqrt{(p_0(1-p_0)/n)}$, assuming one assumption here is that the value of np and $n(1-p)$ should be greater than or equal to 5 because this is this follow binomial distribution. If you if you want to approximate binomial distribution to the normal distribution. So, the assumption is we're np and $n(1-p)$ should be greater than or equal to 5

(Refer Slide Time: 12:48)

Tests About a Population Proportion

Rejection Rule: p -Value Approach

Reject H_0 if p -value $\leq \alpha$

Rejection Rule: Critical Value Approach

$H_0: p \leq p_0$ Reject H_0 if $z \geq z_\alpha$
 $\cancel{p > p_0}$

$H_0: p \geq p_0$ Reject H_0 if $z \leq -z_\alpha$
 $\cancel{p < p_0}$

$H_0: p = p_0$ Reject H_0 if $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$
 $\cancel{p \neq p_0}$



12

What is the rejection rule the same rejection rule if the p -value is less than or equal to alpha we have to reject it, for the critical value approach the same thing if it is Z this is less than or equal

to so p is greater than equal to p_0 this is a right tailed test so this is p less than equal to p_0 left-tail test. So, there right tailed test the Z value is more than $Z \alpha$, reject it. For the left tailed test if the Z value is less than minus α , reject it.

When it is a two-tailed test p equal not equal to p_0 so whether it lies on either side of the minus α by 2 and plus α by 2 we have to reject it.

(Refer Slide Time: 13:26)

Two-Tailed Test About a Population Proportion

Example: City Traffic Police

For a New Year's week, the City Traffic Police claimed that 50% of the accidents would be caused by drunk driving.

A sample of 120 accidents showed that 67 were caused by drunk driving. Use these data to test the Traffic Police's claim with $\alpha = .05$.



13

We well take an example suited traffic police that is the example for the New Year's week the city traffic police claimed that 50% of the accident would be caused by drunk driving. So, the sample of 120 accident showed that 67 where caused by drunk driving, use the data to test the traffic police claimed that alpha equal to 0.05. Similar to that here also the sample data is given and population proportion is given.

What are the sample data first you will solve this p-value approach what are the sample data is given we go back see n is 120 okay the probability actually here the success is 67. So, we have to find out p bar p bar equal to 67 by 120 this is a sample data even. So, the population proportion capital P equal to 0.5 and alpha equal to 0.05. So, now we will use this data we will test the claim that the 50% of accident Road caused by drunk driving.

(Refer Slide Time: 14:49)

Two-Tailed Test About a Population Proportion

$H_0: p = .5$

1. Determine the hypotheses.

$H_a: p \neq .5$

2. Specify the level of significance. $\alpha = .05$

3. Compute the value of the test statistic.

$$\sigma_{\bar{p}} = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{.5(1-.5)}{120}} = .045644$$

$$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} = \frac{(67/120) - .5}{.045644} = 1.28$$

15

So, first we will go for p-value approach the first step in the hypothesis testing is determine in the hypotheses. So, what is the null hypothesis $H_0: P = 0.5$, alternative hypothesis: $P \neq 0.5$ this is given in the problem. The next step is specifying the level of significance alpha equal to 5%. The third step is compute the value of the test statistics that is the value of Z, so for finding the value of Z we need to know the standard error of the population proportion.

So $\sigma_{\bar{p}} = \sqrt{(p_0(1-p_0)/n)}$ the P_0 it is nothing but the assumed population proportion that is 0.5, $1 - 0.5$, n is the sample size 120 so the standard error of the population proportion is 0.045644. Now we will use our traditional said formula, $Z = (\bar{p} - P_0) / \sigma_{\bar{p}}$ here the \bar{p} is nothing but our sample proportion. So, what is the sample proportion out of 120, 67 accidents due to drunk driving.

So \bar{p} is 67 divided by 120 minus this is our assumed a population proportion that is 0.5 the Sigma P already we got it 0.045644 so the Z value is 1.28. when alpha equal to 0.05 because it is a two-tailed test why we are calling it a two-tailed test when you look at the sign of over alternative hypothesis it is not equal to if it is not equal to it is a two-tailed test if it is greater than that is a right tailed test if it is less then, then it is a left tailed test.

Now since the sign of our alternative hypothesis is not equal to type it is a two-tailed test. So, when a alpha equal to 0.05 when you divide by 2 so, this side area is 0.025 this side area is 0.025

when it is a point zero to five the corresponding Z value is 1.96 left hand side is - 1.96 now what has happened our calculated Z value is 1.28. so, 1.28 will be approximately here, 1.28 now when you compare 1.96 and 1.28 this 1.28 is lying on the acceptance region, so we have to accept null hypothesis.

But the methodology for testing the hypothesis what we are using is the p-value approach so the decision of accepting or rejecting our hypothesis not with respect to 1.96 and 1.28. Here we are going to compare the probability so what is that probability is see this left hand side 0.025 the right hand side also 0.025, so what we are going to do when the calculated Z value is 1.28 we are going to look at what is this side area.

Similarly because it is a two tail test when it is - 1.28 we are going to look at this side area by adding this side area plus this side area then after adding the two side area if it is exceeding 0.05 we are going to accept the null hypothesis otherwise we are going to reject it. Now what has happened when it is z-values 1.28.

(Refer Slide Time: 18:54)

The corresponding area is the corresponding area is it is 1.9887 as I told you the fourth state fourth step is compute the p-value so when Z equal to 1.28 the cumulative probability from left to right 0.8997 so when we go back here so this side area when Z values 1.28 this side area is

0.8997 so the right side area this side area will be $1 - 0.8997$ that is approximately 0.1, so what happened but here the area is 0.025.

Now we have to take the decision you can compare 0.025 and 0.1 when they compared 0.1 and 0.25 so the 0.1 is lying on the acceptance side. So, we are about to accept null hypothesis that is the one way to take the decision otherwise the right side area is 0.1 similarly when Z value is -1.28 this left side area is 0.1 so when , when you add this $0.1 + 0.1$ that will be 0.2006 so that value is greater than ever 0.05. So, this added value is greater than 0.05 so we have to accept our null hypothesis.

So, what is the simple rule if the p-value the p-value is less than alpha reject a null hypothesis less than or equal to if the p-value is less than or equal to alpha reject null hypothesis you the p-value is greater than alpha except a null hypothesis. So, now here the p-value that is 0.2006 is more than this so this is the condition for rejection what is the condition the p-value is less than alpha reject it, if the p-value is greater than alpha accept it.

Now what happened now is the p-value is the second condition is satisfied that is the p-value 0.2006 that value is greater than 0.05. So, we are accepting our null hypothesis now we will go back to the step compute the p-value for Z equal to 1.28 the cumulative probability is 0.8997 then the p-value is nothing but $1 - 0.8997$ because the two-tailed test so material by two outer multiplying that you are getting 0.2006 that is greater than 0.05.

So, next we go to next step determine whether to reject H_0 or not because the p-value this 0.2006 is greater than alpha that is a 0.05 we cannot reject it that means we have to accept null hypothesis.

(Refer Slide Time: 22:07)

```

In [13]: from statsmodels.stats.proportion import proportions_ztest
In [14]: count=67
In [16]: samplesize = 120
In [17]: P=0.5
In [18]: proportions_ztest(count, samplesize,P)
Out[18]: (1.28680673975111, 0.1981616572238455)

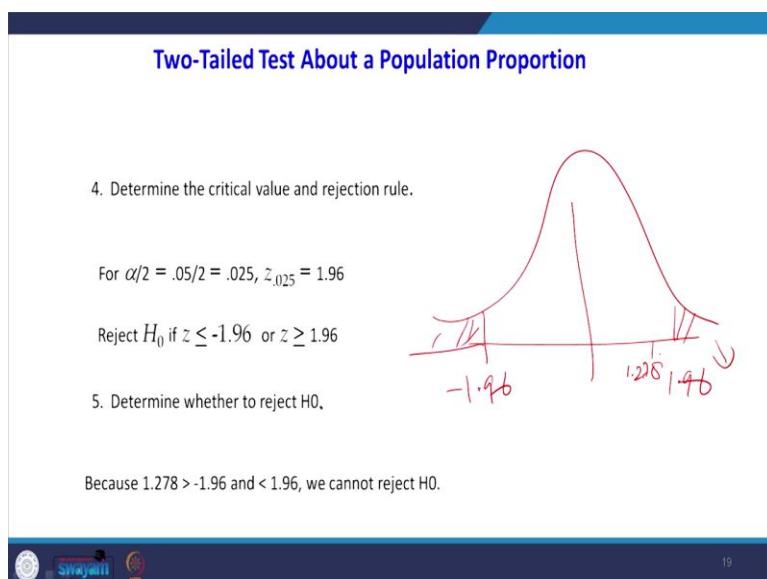
```

17

In Python there is an inbuilt function is there so what to do for that from statsmodel dot stats dot proportion import proportions underscore Z test the proportion test always the Z test there would not be proportion t-test that is number that is not available. So, only if the proportion test means the Z test. So, count is equal to 67 that is the number of success. The sample size is 120, so this capital P is the population mean what we assume it is 0.5, so proportion underscore Z test the syntax is count come on sample size comma capital P.

So, we will get the Z value is 1.28 and the p value is 0.19 so 0.19 because we previous slides at 1.20 it is the address of approximation.

(Refer Slide Time: 23:02)



19

The same proportion test will solve the help of critical value approach determine the critical value and rejection rule because for when it is alpha by 2 that is the for area 0.25 they said well is 1.96 and the left side is - 1.96 if the calculated it is this way 1.96 - 1.96 if the calculated Z value is going this side or this side we are to reject it. So, what has happened the Z value is 1.278 so 1.278 is in the will be here 1.278, so you have to accept the null hypothesis we have to accept the null hypothesis.

Dear students previously we have solved this population proportion test with help of p-value approach then we have solved with the help of critical value approach both a time we have accept a null hypothesis that is the P equal to 0.5. That is a 50% of accident is due to drunk driving. So, I will conclude that in this lecture what you have seen first you have seen t test when you will go for t test when Sigma is not known when the sample size is below 30 we should go for t test.

We have solved one problem for hypothesis testing, we have solved with the help of p-value approach and critical value approach. After that we have solved a problem using a population proportion mean so that means the population proportion is given we have tested whether population proportion can be accepted or not accepted, thank you very much. In the next class we will see different types of error while doing hypothesis testing, thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 19
Errors in Hypothesis Testing

Welcome students in the last lecture we have seen hypothesis formulation and testing. In that hypothesis formulation and testing we have seen what is the null hypothesis and what alternative hypothesis just some introduction about the errors in hypothesis testing. Then we have seen Z test when Sigma is known when Sigma is not known we have done a t-test also.

(Refer Slide Time: 00:57)

Example

- We are interested in burning rate of a solid propellant used to power aircrew escape systems
- Burning rate is a random variable that can be described by a probability distribution
- Suppose our interest focus on mean burning rate
- $H_0: \mu = 50$ centimeters per second
- $H_1: \mu \neq 50$ centimeters per second

Reference: Applied statistics and probability for engineers, Douglas C. Montgomery, George C. Runger, John Wiley & Sons, 2007



2

In this lecture we will go in detail about errors in hypothesis testing we will take an example this example is taken from this book a famous book applied statistics and probability for engineers Douglas C Montgomery and at all it is very interesting books I will like to recommend this book for further reading after seeing this lecture. We are interested in burning rate of a solid propellant used to power air crew escape system.

Burning rate is a random variable that can be described by a probability distribution. Suppose our interest focus on mean burning rate so null hypothesis is μ equal to 50 centimeters per second alternative hypothesis is μ not equal to 50 centimeters per second.

(Refer Slide Time: 01:46)

Value of the null hypothesis

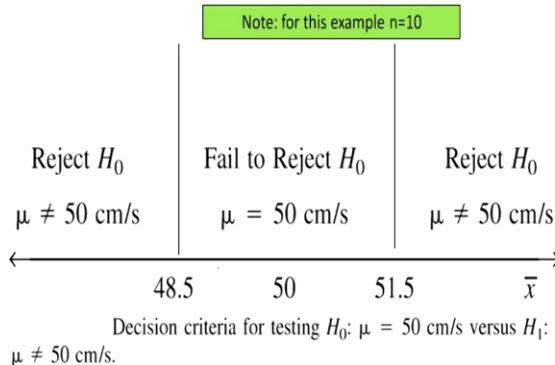
- The value of the null hypothesis can be obtained by
 - Past experience or knowledge of the process, or even from the previous tests or experiments
 - From some theory or model regarding the process under study
 - From external consideration, such as design or engineering specifications, or from contractual obligations



3

You see that in the previous slides we have assumed μ equal to 50 but what is the basis for that 50. There are different possibilities how we can assume the value for our null hypothesis one thing is the past experience our knowledge of the process or even from the previous test our experiments. Other possibility is from some theory or model regarding process under study even from external consideration such as design and engineering specifications are from contractual obligations we can assume the value of null hypothesis.

(Refer Slide Time: 02:23)



4

Suppose I know the say I am following the say confidence interval method. So, I have found the lower limit is 48.5 upper limit is 51.5 when μ equal to 50 centimeters per second. If any mean value which is beyond this 51.5 which is beyond this 51.5 if the sample mean value is beyond

51.5 I will reject null hypothesis the same time if it is below less than 48.5 again I will reject null hypothesis. So, what is happening decision criteria for testing $H_0: \mu = 50$ centimeter per second versus $H_1: \mu \neq 50$ centimeter per second. So in this example we have taken the sample size is 10 and the population standard deviation is 2.5.

(Refer Slide Time: 03:27)

The slide has a dark blue header and footer. The title 'Type I Error' is centered at the top in white. The main content area contains a list of bullet points in black text:

- The true mean burning rate of the propellant could be equal to 50 centimeters per second
- However randomly selected propellant specimens that are tested, we could observe a value of test statistics \bar{x} that falls into the critical region(rejection region).
- We would then reject the null hypothesis H_0 in favor of the alternate H_1 , in fact, H_0 is really true
- This type of wrong conclusion is called a type I error

The footer features the Swastik logo and the number 5.

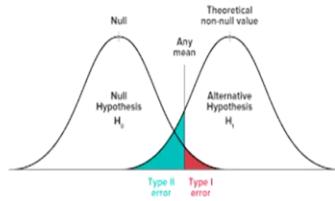
So, by using the previous example I will explain what is the meaning of this type 1 error. See the true mean burning rate of the propellant could be equal to 50 centimeter per second however randomly selected the propellant specimens that are tested we could observe a value of test statistics \bar{x} -bar that falls into the critical region a rejection region. What in the rejection region? Below the lower limit above the upper limit.

So, if our \bar{x} -bar is lying on the rejection region we will reject null hypothesis if the \bar{x} -bar is lying on the acceptance region will accept null hypothesis. So, we will go to the next point we would then reject null hypothesis H_0 if in favor of alternative H_1 in fact H_0 is really true. So, this type of wrong conclusion is called type 1 error. What is the meaning of type 1 error? Even though the null hypothesis are true but the X bar value lying on the rejection side we have rejected null hypothesis this is called type 1 error or incorrect rejection.

(Refer Slide Time: 04:32)

Type I Error

- Rejecting the null hypothesis H_0 when it is true is defined as a type I error



You see for example in this slide you see that there are two normal distribution one is on the left hand side another one is right hand side. So, what will happen this is a one tailed test for this value which is in the red portions sometime the value of x bar may lie on the right hand side that is in the regions. We will reject null hypothesis even though the μ value is same. For example in this case μ value is 50 for example this normal distribution the mean equals to 50.

So, what is the meaning of type 1 error rejecting the null hypothesis H_0 when it is true is defined as the type 1 error. So, what will happen actually the null hypothesis is true because the sample was randomly selected the value of x bar is falling on the rejection region we have rejected a null hypothesis but it is not correct, so, it is incorrect rejection so this is called type 1 error.

(Refer Slide Time: 05:31)

Type II Error

- Now suppose the true mean burning rate is different from 50 centimeters per second, yet the sample mean \bar{x} falls in the acceptance region
- In this case we would fail to reject H_0 when it is false
- This type of wrong conclusion is called a type II error



7

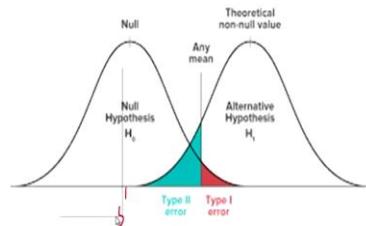
Now we will explain what is the meaning of type 2 error?. Now suppose the true mean burning rate is different from 50 centimeter per second yet the sample mean x -bar fall in the acceptance region what will happen this, now the null hypothesis is not true but the sample mean is following in the accepting region in this case we would fail to reject H_0 , when it is false. So, what is happening the H_0 is false but still we have accepted our H_0 , so it is false acceptance so this type of wrong conclusion is called type 2 error.

So I just am saying what do you need I type 1 type 2 error see the type 1 error is incorrect rejection type 2 error is false acceptance.

(Refer Slide Time: 06:22)

Type II Error

- Failing to reject the null hypothesis when it is false is defined as a type II error



8

You see this situation there are two normal distribution one is we have seen this is 50 say it is 52 now what has happened there are 2 population which are overlapped my concern is about the population whose mean is 50. But there is another population whose mean is 52 which are overlapping. So, what will happen this much regions is not belongs to that light green one it is not belongs to this population. So, this region this region is belongs to population whose mean is 52 just because off it is lying on acceptance side of this population where the mean is 50.

We are wrongly we are falsely accepted, so it is called a type 2 error, so, failing to reject null hypothesis when it is a fall is defined as the type 2 error.

(Refer Slide Time: 07:06)

Type I and Type II Errors		
	H_0 is correct	H_0 is incorrect
H_0 is accepted	correct decision	Type II error (β) Incorrect acceptance
H_0 is rejected	Type I error (α) Incorrect rejection	correct decision

You see the type 1 type 2 error see the H_0 is correct but we have rejected. So, this is your incorrect rejection alpha, the H_0 is incorrect but we accepted this is a type 1 error. So, that is a false acceptance this we have seen previously also the same table.

(Refer Slide Time: 07:28)

Type I error

- In the propellant burning rate example, a type I error will occur when either $\bar{x} > 51.5$ or $\bar{x} < 48.5$ when the true mean burning rate is $\mu = 50$ centimeters per second
- Suppose the standard deviation of burning rate is $\sigma = 2.5$ centimeters per second and $n = 10$
- Probability distribution $\mu = 50$, standard error = 0.79.
- Type I error is

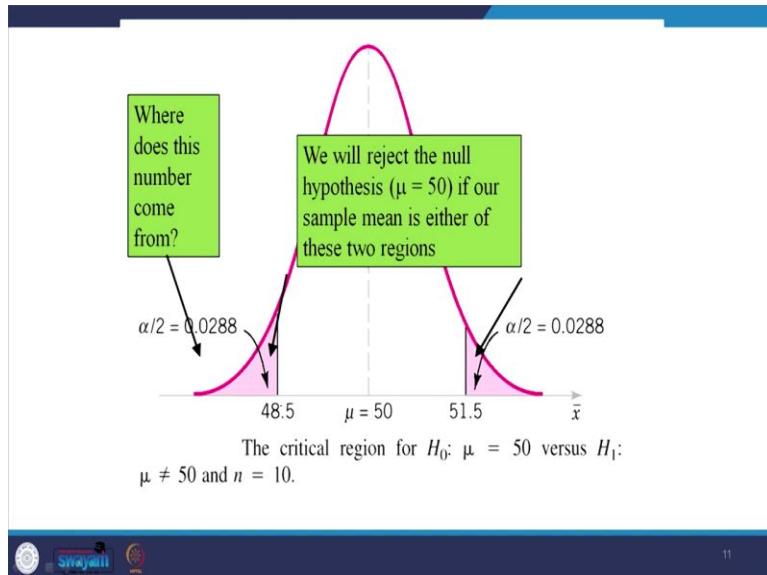
$$\alpha = P(\bar{x} < 48.5 \text{ when } \mu = 50) + P(\bar{x} > 51.5 \text{ when } \mu = 50)$$

10

Now you will see how to calculate type 1 error and type 2 error first we will go with the type 1 error. In the propellant burning rate example a type 1 error will occur when either the sample mean is greater than 51.5 or sample mean is less than 48.5 when the true mean burning rate is mu equal to 50 centimeters per second. Suppose the standard deviation of the burning rate is that is a sigma is 2.5 centimeters per second and n equal to 10, so the probability distribution mu equal to 50 the standard error actually standard error is our Sigma by root n that value standard error is 0.759.

So what is the value of the type 1 error is when the probability of $X \bar{x}$ less than or equal to 48.5 when true mean is 50 plus when the $X \bar{x}$ is greater than 51.5 when true mean equal to 50.

(Refer Slide Time: 08:26)



You see that you see in the left hand side it is a 48.5 when the sample mean is what is the probability of sample mean to lie below 48.5 plus what is the probability of that sample mean to lie above 51.5 when you add that that corresponding probability is nothing but your type 1 error. You see that we got alpha by 2,= 0.0288 similarly on the right hand side you will see how this has come.

(Refer Slide Time: 08:55)

```
Defing function for calculating alpha value

In [6]: def z_value(x,mu,SEM):
    z = (x - mu)/SEM
    if(z < 0):
        alfa = stats.norm.cdf(z)
    else:
        alfa = 1 - stats.norm.cdf(z)
    print (alfa)

calculating alpha for different values of x,mu, and SEM

In [8]: x =48.5
        mu = 50
        SEM = 0.79

In [9]: z_value(x,mu,SEM)
        0.02879971774715278
```

Using Python actually I have pasted the print screen of the Python after running first we for finding the type 1 error will define a function. So, that function I'm going to call it as def of Z underscore value X, mu, sem standard error of mean you see that whenever we define function there should be a colon. So, first I am finding Z value Z value is x - mu by standard error, if the Z

value is less than 0 that means if the value of it will be like this if the Z value is the Z values on the negative side simply the p value is nothing but the cumulative distribution function of Z.

So, when you type alpha equal to alpha I just I am naming equal to stat.norm.cdf Z you will get alpha value suppose if the Z values greater than 0 we have to find out the right side area so if we want to know the right side area you the whole area has to be subtracted from one so that we will get the right side area it, else : alpha equal to 1 – stats.norm.cdf Z so print alpha, so calculating alpha for different value of x mu and standard error of mean.

You see first I will find out the left side area so when area suppose this is X values 48.5 what will be the area? So, X is 48.5, mu this is 0 for standard normal distribution but we at present we are taking mu equal to 50 because after converting to Z scale to become 0, standard error of the mean is Sigma by root n 0.75. Now we will call that a function which we are defined. So, Z value so we have to give the value of x because that function is defined 48.5 mu is 50 standards of the mean so the left side area will get 0.0287 this value is given left side also see that 0.0288 right.

Now we have to find out the right side actually this is alpha by 2, value whatever value which we got it this value is alpha by 2. If you put the upper value that is 51.5 alpha by 2, so if you replace instead of 48.5 if you if you put 51.5 here when you replace this X and when you replace 51.5 what will happen this Z value will be positive the Z value is positive this command will be executed. So, if you want firstly they will find this left side area then from one the left side area will be subtracted then we will get the right side area so we will get another 0.0287.

(Refer Slide Time: 12:00)

Type I error

- Type I error = 0.057434
- This implies that 5.7 % of all random samples would lead to rejection of the hypothesis $H_0: \mu=50$ centimeters per second.
- We can reduce the type I error by widening the acceptance region. If we make critical value 48 and 52, the value of alpha is 0.0114 (adding 0.0057 and 0.0057).
- Change sample size to 16 then alpha is 0.0164.

In [40]: `z_value(48,mu,SEM)`

0.005676434117424844

In [41]: `z_value(52,mu,SEM)`

0.0056764341174248



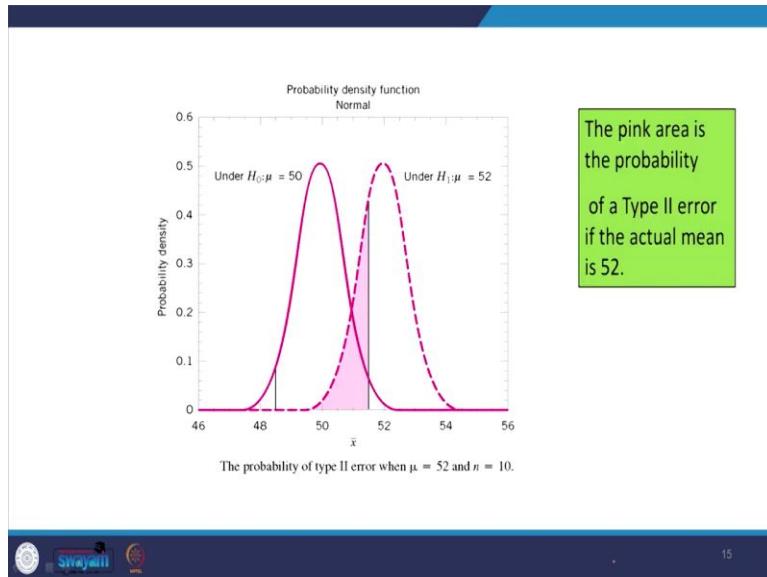
13

When you add this $0.0280 + 0.0287$ you will get the Alpha value so that alpha is 0.057 this is your value of alpha error type 1 error. So, what is the meaning of this type one error this implies that 5.7% of all random samples would lead to rejection of null hypothesis that is H_0 when μ equal to 50 centimeters per second. So, there is a possibility of rejecting to null hypothesis is 5.7%, so we can reduce the type one error 1 possibility is by widening the acceptance region.

What is the widening the accept region if you make critical value 48 and 52 what will happen what is the widening of this acceptance region is suppose this is currently this one so when you increase now it is 51.5, so now you make it this right hand side 52 left hand side 48 so what is happening the acceptance region is widened yeah you see that lower side on 48 upper side 52 so what is the area cutting $0.00567 + 0.0567$ when you add it will become 0.0114.

So, what is happening the type 1 error can be reduced by increasing the acceptance region that is one possibility another possibility is if you increase the sample size the previous problem you have taken sample size is 10, now from 10 if you increase 16 what is happening the value of alpha is decreasing that means the Alpha is decreasing means we are more accurate in making decision that is the possibility of incorrectly rejection is reduced.

(Refer Slide Time: 13:51)



Next we will go to type 2 error first we will explain what is a type 2 error and I will take 2 example 2, 3 examples to calculate the value of type 2 error. In the previous example what has happened there are 2 population which are overlapped so this is 50 this is 52 the rejection region this is this much when the population mean is 15 there is another population whose mean is 52 which is over lapping.

So, this overlapping region is that pink one because the pink portion is lying on acceptance side we have falsely accepted assuming that that region has come from the population mean whose value is 50. So, it is a false acceptance right so this pink region is nothing but the value of your type 2 error.

(Refer Slide Time: 14:45)

Type II Error

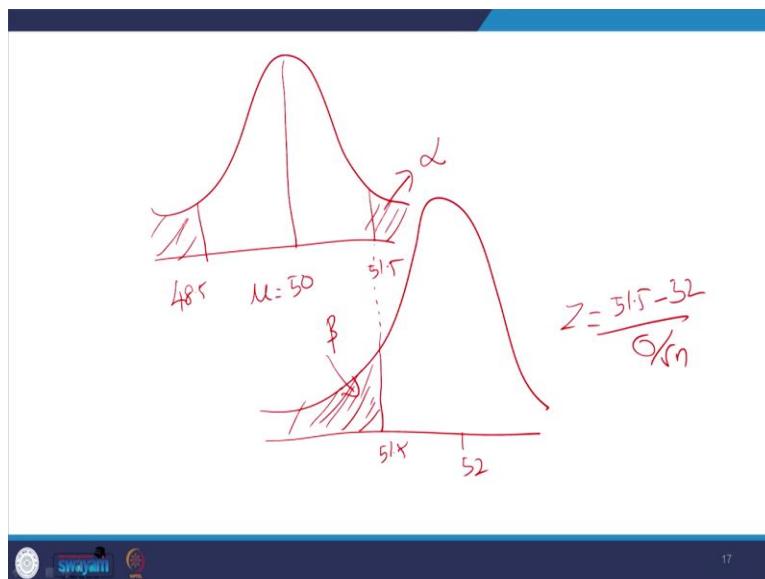
- Type II error will be committed if the sample mean \bar{x} falls between 48.5 and 51.5 (critical region boundaries) when $\mu = 52$. $\beta = P(48.5 \leq \bar{x} \leq 51.5 \text{ when } \mu = 52)$
- 0.2643
- When $\mu = 50.5$
- 0.8923

```
In [4]: beta = stats.norm.cdf((51.5-52)/0.79) #
In [5]: beta
Out[5]: 0.26339575390741593
In [8]: beta = stats.norm.cdf((51.5-50.5)/0.79) #
In [9]: beta
Out[9]: 0.8972117321157791
```

The probability of type II error when $\mu = 52$ and $\sigma = 10$.

Suppose how to find out this type 2 error see type 2 error will be committed if the sample mean \bar{x} falls between 48.5 and 51.5 critical region when μ equal to 52 you see that it is a 52 not μ equal to 50 when I say 52 that population is nothing to do when the mean equal to 52 population because it is something other population but it is lying on the acceptance side so I am accepting falsely I have accepted so that is our type 2 error. So, what is the possibility so this is worth 51.5.

(Refer Slide Time: 15:20)



So what is happening there are 2 population is overlapped. So, this is our original μ equal to 50 so this region is 51.5 this left hand side is what is that value, 48.5 this is for my assumed hypothesis value in μ equal to 50 but actually what has happened there is another population

whose mean is 52, so I am extending this so this much portion this much portion is not really belongs to the first one, but lying on the acceptance side so this much portions I have falsely accepted so this portion is called the beta this portion is called alpha.

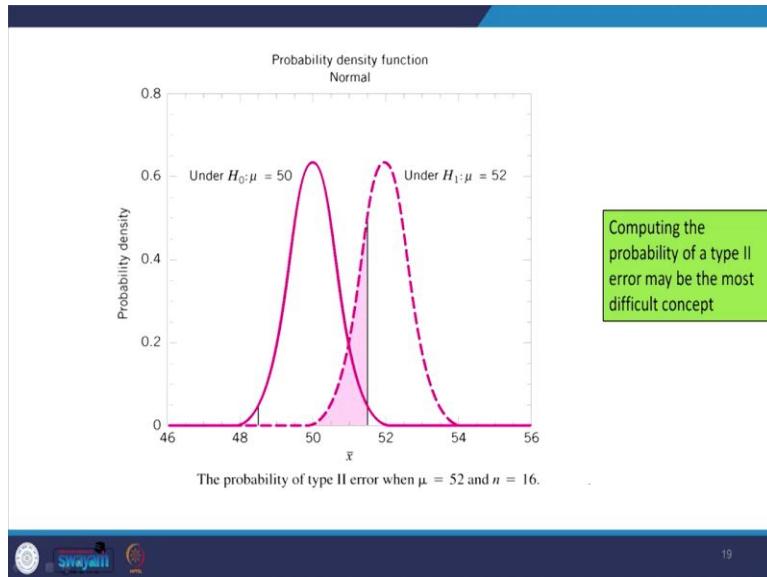
So we have to find out the value of beta and one more thing the value of alpha plus beta \neq 1, it should be very careful. So, what is happening this is 51.5 so this also 51.5 because this condition is same. Now for this population that is which is on the below we have to find out the left side area then we how do you do that first you have to find out the Z value Z value is $(X - \mu)$ by Sigma, so $51.5 - 52$ divided by Sigma by root n.

So, what will happen for this region when Z it is below the left hand side the corresponding value is your type 2 error. See, that I have done that also so I calling it the beta stats.norm.cdf cumulative distribution function. So, my x value is 51.5 minus my true mean is 52 divided by Sigma by root n so 0.7 so that beta value is 0.263. So, this 0.263 is nothing but this value 0.263 is 0.264, so this is that pink region area is this one just I will give an explanation the code over so Python is given how to find out.

Now you see that no μ equal to 50.5 what happened again between the true mean and assumed mean the difference is becoming less. So, if I put the true mean is 50.5 you see that I have changed there 52 here I have changed in 50.5 so again beta equal to stats.norm.cdf of 51.5 there is nothing but $X_{\bar{}} - \mu$ by Sigma by root n that value is 89 you see that the value of beta has increased. So, one point at present you have to remember when the difference is decreasing between your assumed mean and the true mean that type two error is increasing.

We can say one example suppose there are two product goodness original another one is duplicate both are looking similar color wise texture wise quality wise, there are more possibility for committing type two error when whenever the difference between original and duplicate of product is very, very less, the same way whenever the mean the published which you assumed and the true mean the difference is less there is more possibility of committing type 2 error that I will explain.

(Refer Slide Time: 19:15)



You see this one again when the mean is 50.5 there is a more there are more pink region that is there is more type 2 error. So, the point is when the distance between your assumed mean and the true mean is decreasing there are more possibility of committing type 2 error. Now we will see computing of this type two error already I have explained so computing the probability of type two error may be the most difficult to concept if you are doing it with the help of statistical table it will take more time that is why we go for Python that it will solve your problem very quickly. So, this area we have found out already.

(Refer Slide Time: 19:57)

acceptance region	sample size	α	β at $\mu = 52$	β at $\mu = 50.5$
$48.5 < \bar{x} < 51.5$	10	0.0576	0.2643	0.8923
$48 < \bar{x} < 52$	10	0.0114	0.5000	0.9705
$48.5 < \bar{x} < 51.5$	16	0.0164	0.2119	0.9445
$48 < \bar{x} < 52$	16	0.0014	0.5000	0.9918

For constant n , increasing the acceptance region (hence decreasing α) increases β .

Increasing n , can decrease both types of errors.

Now look at this table acceptance region is their sample size is there you see that for the same sample size when you increase the acceptance region what is happening the Alpha is decreasing,

sample size is same you are widening the acceptance region, so the Alpha is decreasing. When alpha is decreasing you see what is happening the beta is increasing. So, the relation between alpha and beta is when you decrease alpha beta will increase it is like this as I told you previously there are two normal distribution this is your rejection region there is another normal distribution this is mu equal to 50 this is 52.

So this side portions, so this is your alpha this side portion is beta I will change the color. So, this portion the green portion is nothing but your type 2 error the red portion is your type on our alpha and beta. So, what is happening suppose assume that I am keeping in that line where there's intersections there I am keeping a pen like this if I move towards right-hand side what will happen alpha will decrease so when alpha is decreasing what is happening to beta, beta value is increasing.

Suppose I am keeping this pen I am moving towards left hand side what will happen beta will decrease but alpha will increase so that is explained with the help of this table so the relation between relationship between alpha and beta is inverse, apart from this equal to the last column the MU equal to 52 again I am the mu through 50.5 suppose the difference is decreasing what is happening the value of beta is increasing that is one point.

Now we look at the another point suppose when you increase the sample size you keep the acceptance region as it is you increase the sample size 10 to 16 you see that when you go to 10 to 16 you see here 48.5, 51.5 sample size is 10 alpha is 0.05 for the same reason when you increase you compare this will go to say I am comparing this accepted region is same but I have increased my sample size the Alpha value is decreasing.

See initially our alpha value 0.05 now it is 0.01 look at this beta also the value of beta also you see initially it is 0.26 now it is 0.21, so what the point we are learning from here is when you keep acceptance region as the constant one when you increase the sample size both value of alpha and beta will decrease okay that is the point here let us see this. For constant n when you increase the acceptance region alpha is decreasing with the Alpha is decreasing beta values increasing.

Second point increasing n can decrease both type of error that is type 1 and type 2 there is a learning from this slide.

(Refer Slide Time: 23:49)

Type I & II Errors Have an Inverse Relationship

If you reduce the probability of one error, the other one increases so that everything else is unchanged.

21

Type 1 and type 2 errors having an inverse relationship if you reduce the probability of one error the other one increases so that everything else unchanged. So, that is a relation between alpha and beta remember alpha + beta is not equal to 1.

(Refer Slide Time: 24:08)

Factors Affecting Type II Error

- True value of population parameter
 - β Increases when the difference between hypothesized parameter and its true value decrease
- Significance level α
 - Increases when β decreases
- Population standard deviation σ
 - Increases when β increases
- Sample size
 - β Increases when n decreases

22

Now let us see a factors affecting type 2 error the true value of population parameter beta increases when the difference between this point already I told you the difference between

hypothesis the parameter and its two values decreasing, significance level alpha increases when beat decreases, population standard deviation Sigma increases when beta increases, sample size beta increases when n decreases that is the relation between your different element of your type 2 error.

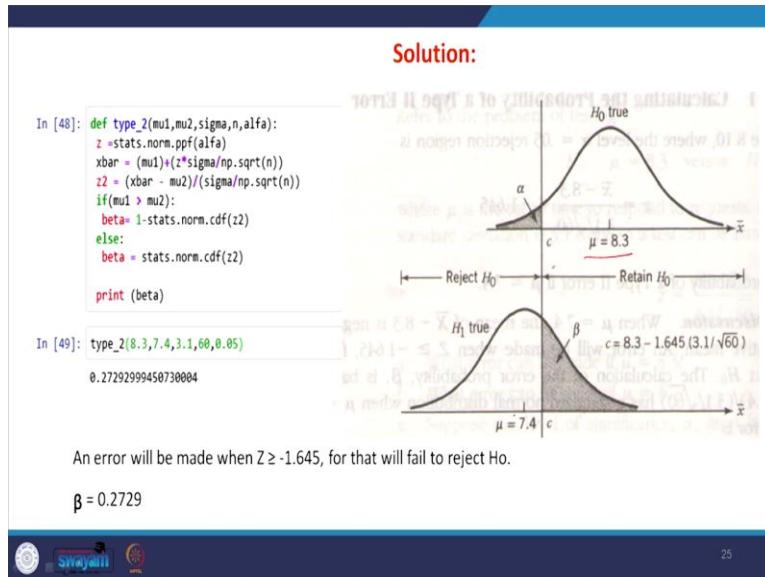
(Refer Slide Time: 24:49)

How to Choose between Type I and Type II Errors

- Choice depends on the cost of the errors
- Choose smaller Type I Error when the cost of rejecting the maintained hypothesis is high
 - A criminal trial: convicting an innocent person
- Choose larger Type I Error when you have an interest in changing the status quo

How to choose there between type 1 and type 2 error the choice depends upon the cost of error. So, the first point is choose smaller type 1 error when the cost of rejecting the maintained hypothesis high, for example in a criminal trial committing an innocent person is very, very costly mistake so that time the value of alpha should be very less. Choosing a larger type 1 error when you have an interest in changing the status quo so what will happen if you are willing to change the status quo if you increase the alpha value obviously there is a more chances for a rejection region either hypothesis get rejected that is the status quo is getting rejected.

(Refer Slide Time: 25:40)



We will take another problem to find out type 2 error so I am assuming $\mu = 8.3$ alternate hypothesis $\mu < 8.3$, it is a left tailed test. Determine the probability of type 2 error if the true mean is 7.4 at 5% significance level when Sigma is 3.1 and n equal to 60. See that I have drawn this one my assumed mean is 8.3 the question is asked if the true mean is 7.4 what is the value of beta? You see that if any portions which are going on left hand side I will reject it so this side I will reject it but the right hand side I will accept it.

But what is happening the true mean is 7.4 it is lying on acceptance side of where the μ equal to 8.3, I actually I have to reject this since it is lying on the acceptance region so this much portions this much portion I have falsely accepted so that beta value is nothing but your type 2 error. See that the value of C is constant for this population and this population so we have to find out this right side area. So, for this purpose I have developed a function because this function is very useful first you let us understand.

So I am going to define your function I am going to call it as a type_2 so what are the parameter which I am going to take mu1 my assumed mean mu2, true mean Sigma population standard deviation n sample size alpha significance level colon. So, for example the first one

which is in the topper 1 the normal distribution and find I'm finding out what is the Z value in this location what is the Z value I know what is alpha value so Z equal to stats.norm.cdf of alpha

value if we substitute what will happen I will get the value of Z. If I know the value of Z I can find out X bar how I can find out because Z equal to in this relationship if I know Z value I can find out X bar. So, X bar is when you mu multiplied by mu plus when you bring this left hand side Sigma Z multiplied by Sigma by root n that is it an X bar equal to mu 1 + Z star Sigma divided by np because from numpy that is kernel numpy .square root of n.

So, I will get X bar this is the value of Z now I have to find out. The Z 2, this said to this what will be this Z 2 this this X bar this X bar this value is X-bar so corresponding normalized scale is Z 2, so what will happen Z 2 is X bar - mu 2 that is for the from this population what is the mean that is a 7.4 X bar - mu 2 divided by Sigma by np . square root n. You see the condition if mu 1 is greater than mu 2 what is mu 1, mu 1 is now 8. 3 this is 7.4 in this case yes mu 1 is greater than mu 2 what will happen I will get the positive Z value if the Z is positive if I want to know the p value from 1 I have to subtracted.

So the Z value is positive the beta equal to 1 - stats.norm.cdf of Z 2 if Z value is negative just finding the left side value beta equal to stats.norm.cdf Z 2, beta. So, when you type this in Python type underscore now we have to give this value of mu 1 suppose what I am going to do a good to find out the beta value the mu 1 is the zoom in 8.3 mu 2 is true mean 7.4 Sigma is 3.1 sample size is 60, alpha equal to 0.05. so, now what will happen here if it is a positive value the corresponding probabilities 0.2729 that is why that value yes this 0.729.

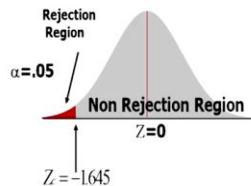
So, the right side beta will is 0.2729 so when will commit to type 2 error and the error will be made when Z values greater than so now is here the corresponding Z value is – 1.645 whenever you crossing - 1.645 on the right hand side then you will accept that that is nothing but your value of type 2 error.

(Refer Slide Time: 30:39)

Solving for Type II Errors: Example

$$H_0: \mu = 12$$

$$H_a: \mu < 12$$



$$\bar{X}_c = \mu + Z_c \frac{\sigma}{\sqrt{n}}$$

$$= 12 + (-1.645) \frac{0.10}{\sqrt{60}}$$

$$= 11.979$$

If $\bar{X} < 11.979$, reject H_0 .

If $\bar{X} \geq 11.979$, do not reject H_0 .

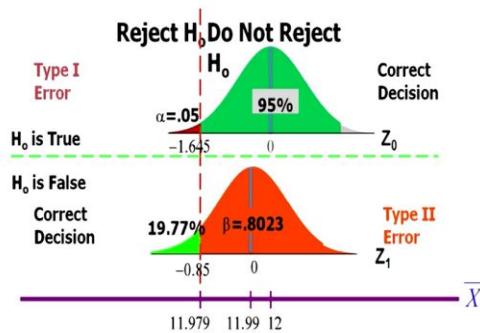


26

I will take another example solve you for type 2 error one more example mu equal to 12, mu less than equal to 12 we know that X bar equal to mu 0 Sigma by root n so here assumed means 12, Z value is because the left side you when alpha equal to 0.05 corresponding Z is - 1.645 Sigma values given the X bar is 11.979 if the value of X bar is below 11.979 we will reject it if the X bar is above 11.979 I will accept my null hypothesis.

(Refer Slide Time: 30:39)

Type II Error for Example with $\mu = 11.99$ Kg

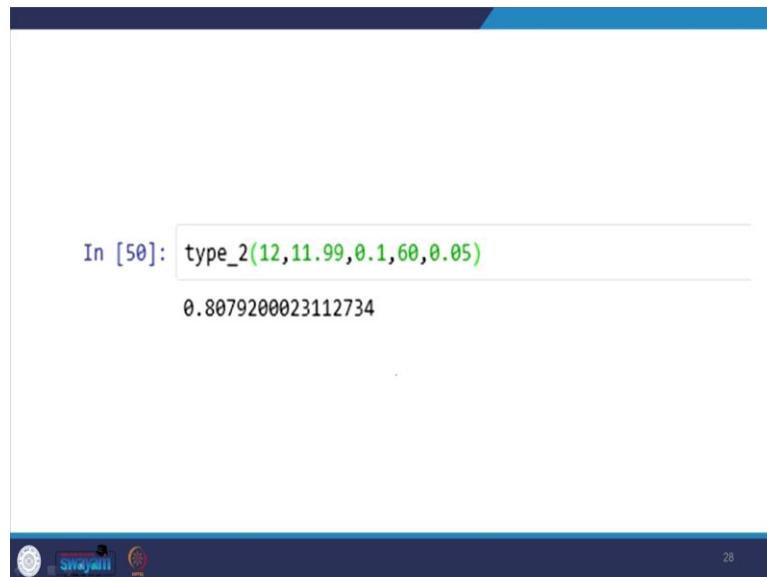


27

Now we look at this graph you see that my assumed mean is 12 suppose my true mean is 11.979 what are you with the value of type 2 error you see that when alpha equal to 0.05 - 1.645 I know that so what will happen if the true mean is 11.979 what will happen Z value will be X bar - mu

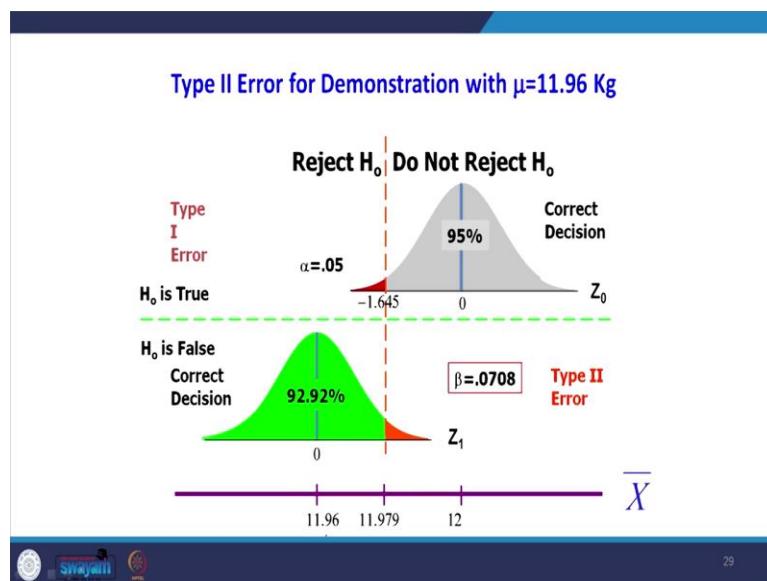
so $(11.979 - 12)/\text{Sigma by root n right}$, we will get Z value so corresponding right hand side area is my type two error. So, beta equal to 0.8023.

(Refer Slide Time: 32:09)



I have done some code for this so type two error so twelve is what is that my assumed mean 11.99 is my true mean this is my standard deviation Sigma value this is my n this is my alpha, so it is 80 that is why this much here, here so what we are getting we are getting 80790 approximately 80%.

(Refer Slide Time: 32:39)



I will go to see another problem the true mean is 11.96 little far away from the 12 what is that we the true mean is going towards left hand side now what has happened to the value of beta, so

this portion is my beta type 2 error. So, by use already we have developed your function for that will substitute this 12, 11.9, 6.1, 60, .05 so what is happening the value of beta is becoming very less. Again I am stressing this point you the difference is bigger the value of beta is very less if the difference is closer the earlier beta is very high.

(Refer Slide Time: 33:19)

The slide has a blue header bar with the title 'Hypothesis Testing and Decision Making'. Below the title is a list of bullet points:

- We have illustrated hypothesis testing applications referred to as significance tests
- In the tests, we compared the p -value to a controlled probability of a Type I error, α , which is called the level of significance for the test
- With a significance test, we control the probability of making the Type I error, but not the Type II error
- We recommended the conclusion "do not reject H_0 " rather than "accept H_0 " because the latter puts us at risk of making a Type II error

At the bottom of the slide, there is a footer bar with icons for a clock, a person, and a globe, followed by the text 'SWAYAM' and 'GOVT OF INDIA'.

Now hypothesis testing and decision making we have illustrated hypothesis and testing applications. Now let us see the; what is the application of this type 2 error we have illustrated hypothesis testing applications referred to as a significance test. In the test we have compared the p -value to a controlled probability of type 1 error alpha which is called the level of significance for the test. What do we have done to accept or reject a null hypothesis we have considered the p value that is compared with alpha.

The p value is smaller than the Alpha we have rejected it the p value is greater than alpha we have accepted it. So, we will go to the next point with the significance test we controlled the probability of making type 1 error but not the type 2 error. We recommended the conclusion do not reject H_0 actually we have to use accept H_0 but very cautiously we have used do not reject H_0 rather than accept H_0 because the later puts us at risk of making type 2 error.

Why we are not accepting there is no proof that the value which have assumed in a null hypothesis correct so that is why we are saying do not reject it now in this example what we are

going to do what should be the value of our null hypothesis. So, that the something called the power of test can be improved we will see the definition of power of test. With the conclusion do not reject H₀ the statistical evidence is considered inconclusive you are not able to say anything.

Usually this is an indication to postpone a decision until the further research and testing is undertaken. But in many decision-making situations the decision-maker may want and in some cases may be forced to take action both the conclusion do not reject or the conclusion reject H₀ in such situation it is recommended that a hypothesis testing processor be extended to include consideration of making type 2 error.

So, what we are going to say what in whenever you do the hypothesis our testing we have to see the possibility of committing type 2 error also that I will show you with the help of an example.

(Refer Slide Time: 35:48)

The slide has a dark blue header and footer. The main title 'Power of a test' is centered at the top in a blue font. Below the title is a bullet-point list of four items. To the right of the list is a cartoon illustration of a meal on a plate, featuring a sandwich with meat and vegetables, and a bowl of red soup with a spoon. The footer contains three small icons: a person, a book, and a gear, followed by the text 'ShreyasIT'.

- The mean response time for a random sample of 40 food-order is 13.25 minutes
- The population standard deviation is believed to be 3.2 minutes.
- The restaurant owner wants to perform a hypothesis test, with $\alpha = 0.05$ level of significance, to determine whether the service goal of 12 minutes or less is being achieved.

That point is called power of test ok, suppose there is a restaurant is there suppose in the restaurant when you order some dosa and you order some coffee or soup many times it will take different times, sometimes they because they were to prepare it. Assume that the owner of the restaurant has the target of service goal of 12 minutes or less whether it can be achieved or not so what is the different possibility of committing type 2 error if you assume mu equal to 12. You will see that the problem detail.

The mean response time for a random sample of 40 food order is say 13.25 minutes the population standard deviation is believed to be 3.2 minutes the restaurant owner wants to perform a hypothesis test with alpha equal to 5% a significance level to determine whether the self-service goal of 12 minutes are less is being achieved. Now what is happening first you have to start null hypothesis so null hypothesis is the status quo.

(Refer Slide Time: 37:00)

Calculating the Probability of a Type II Error

Hypotheses are: $H_0: \mu \leq 12$ and $H_a: \mu > 12$

Rejection rule is: Reject H_0 if $z \geq 1.645$

Value of the sample mean that identifies the rejection region:

$$z = \frac{\bar{x} - 12}{3.2/\sqrt{40}} \geq 1.645$$

$$\bar{x} \geq 12 + 1.645 \left(\frac{3.2}{\sqrt{40}} \right) = 12.8323$$

We will accept H_0 when $x \leq 12.8323$

34

The status quo is $\mu \leq 12$ alternative hypothesis is $\mu > 12$ so what will happen what kind of test this is this is right tailed test. So, this value is 12 if anything goes this side I will reject it so when alpha equal to 0.05, so the corresponding value is 1.645 if any value Z value goes beyond 1.645 I will reject it. So, we will substitute this value into our Z formula so Z equal to $(\bar{X} - 12) / (3.2/\sqrt{40})$, if it is greater than 1.645 I will reject it.

So from this relationship I will bring the value of \bar{X} okay we are finding the value of \bar{X} that is a 12.83 so what will happen we will accept H_0 when the value of \bar{X} is 12.83, so this value will 12.83 if anything value goes that side we will accept H_0 if anything goes below this will reject it.

(Refer Slide Time: 38:05)

Calculating the Probability of a Type II Error

Probabilities that the sample mean will be in the acceptance region:

Values of μ	$z = \frac{12.8323 - \mu}{3.2/\sqrt{40}}$	β	$1-\beta$
14.0	-2.31	.0104	.9896
13.6	-1.52	.0643	.9357
13.2	-0.73	.2327	.7673
12.8323	0.00	.5000	.5000
12.8	0.06	.5239	.4761
12.4	0.85	.8023	.1977
12.0001	1.645	.9500	.0500



35

So, what will happen here in this we assume you said assumed we have assumed $\mu \leq 12$, $H_1: \mu > 12$ now the question is how what is the logic behind this 12. So what I am going to do instead of this 12, I am going to supply different values of this new value say I'm going to supply 14 13.6 13.2 12.00 so value of μ so Z it is $(12.83 - \mu) / (3.2/\sqrt{40})$, in this Z formula when you substitute the value of μ 14 this is the Z value.

So when the Z value is -2.31 what is the value of type two error 0.01 so $1 - \beta$ so this $1 - \beta$ is nothing but power of a test. Power of a test is rejecting a null hypothesis when it should be rejected. So now instead of 14 if I make 13 so again the Z value is -1.52, so corresponding β will be 0.64 you see that when the difference is becoming closer to 12, what is happening you see that the value of β is increasing whenever value of β is increasing the power of test is decreasing.

(Refer Slide Time: 39:43)

```

In [20]: type_2(14,12,3.2,40,0.05)
0.010499750448532241

In [21]: type_2(13.6,12,3.2,40,0.05)
0.06457982995225997

In [23]: type_2(13.2,12,3.2,40,0.05)
0.233657510104159

In [22]: type_2(12.8323,12,3.2,40,0.05)
0.49995865746353273

In [27]: type_2(12.8,12,3.2,40,0.05)
0.5254013387545549

In [24]: type_2(12.4,12,3.2,40,0.05)
0.8035262335707292

In [26]: type_2(12.0001,12,3.2,40,0.05)
0.9499796127157129

```



36

So, how we got this 0.0104 I have done in the next, next slide see I am calling that function which I previously used so type_2 if true mean is 14 assumed mean is 12 Sigma is 3.2 n equal to 40 alpha equal 2.05, so my beta is 0.01 this is for my this is 14. Suppose if it is a 13.6 right substituting 13.6 or other value my beta well is 0.06, so that is this value so when it is 13.2 what is the beta value so in substituting 13.2 the beta value is 0.23 and 0.23. If it is 12.8 again beta value is 0.5 and so on.

(Refer Slide Time: 40:35)

Power of the Test

- The probability of correctly rejecting H_0 when it is false is called the power of the test.
- For any particular value of m , the power is $1 - b$.
- We can show graphically the power associated with each value of μ ; such a graph is called a power curve.

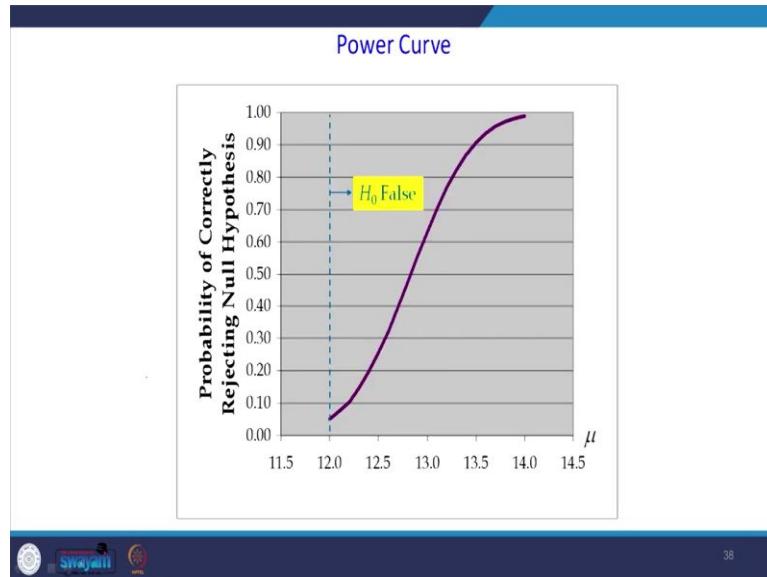


37

So, if I plot this okay now we will come to before plotting I will define what is this power of test the probability of correctly rejecting null hypothesis when it is false is called the power of test. For any particular value of mu this is mu the power is 1 - beta you call it is capital B that is

convenient. So, we can show graphically the power associated with each value of mu such graph is called power curve.

(Refer Slide Time: 41:08)



So, the application of power curve first I will explain what is the element in this power curve. Here in the x-axis the true mean which I have assumed in the y-axis the value of 1 - beta probability of currently rejecting null hypothesis what is happening when the difference is increasing between true mean even your assumed mean there is a more chances you will correctly reject your null hypothesis. So, this power of test says what should with the value of mu you see that when you feel if he assume you equal to 12 power is less when the power is when you assume you equal to 14.5 this power is more.

So this power curve is helping us to decide what is the possibility of committing type 2 error at the same time how much value of null hypothesis we can have so that we can improve our power of a test otherwise we can decrease the beta. Dear students in this lecture we have seen different types of error while doing hypothesis testing. We have taken one practical example I have explained what is the meaning of type 1 error and type 2 error.

And also we have calculated value of type 1 error and type 2 error at the end we have seen a power of a hypothesis testing we call it as a power curve. So, what we have done we have suggested what is the possible value of mu and corresponding beta value are a corresponding

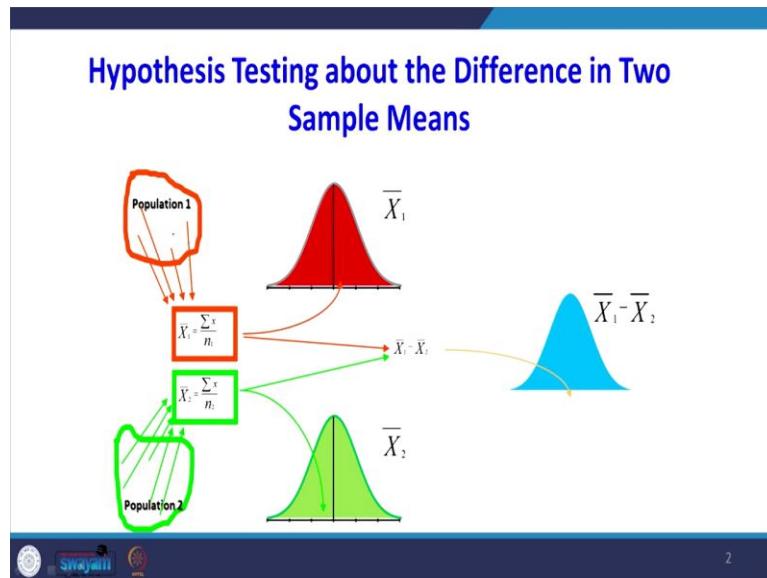
power of a test. So, with that we will conclude in this lecture in the next lecture we will go for a two sample hypothesis testing, thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 20
Hypothesis Testing: Two Sample Test-I

Dear students today we are entering into another topic that is a hypothesis testing for 2 sample tests. First I will explain what is the theory behind this 2 sample test?

(Refer Slide Time: 00:38)



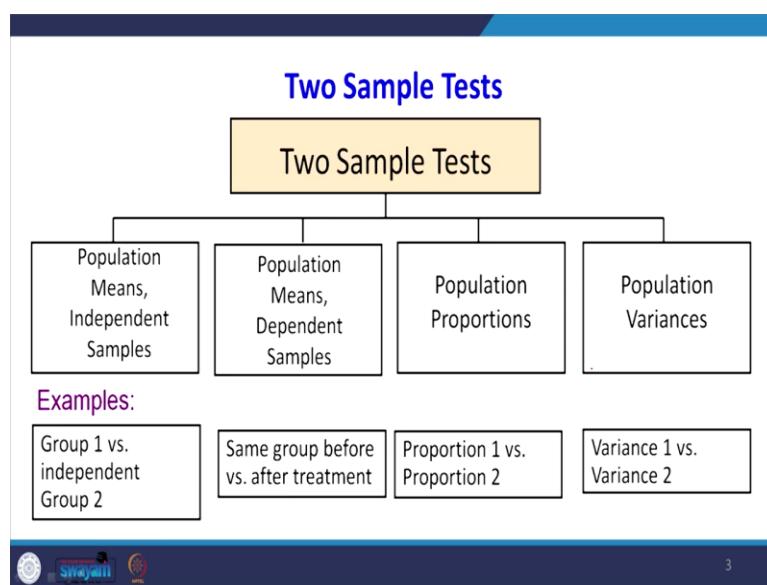
You look at this picture see there are 2 population there are population 1 and population 2 suppose if I take some sample from population 1 I am taking sample and finding sample mean I am calling it as X_1 bar that is a Sigma X divided by n_1 there is another population which is in green color from that I am taking some sample X_2 then finding mean of that sample by using this formula Sigma X_2 divided by n_2 .

If I plot this is the sampling distribution of X_1 bar the green one is the sampling distribution of X_2 bar. So, the mean of this sampling distribution of X_1 bar is μ_1 we can say μ_1 the mean of sampling distribution of X_2 is μ_2 . The variance is σ_1^2 divided by n_1 now what will happen because this result is so the variance actually you know to right this way variance of X_1 bar, this is variance of X_2 bar.

Suppose if I find the difference of their sample mean if I plot the difference that will follow a normal distribution. So, the mean of this population is $\mu_1 - \mu_2$ the variance of this population is so the variance is we have to find the difference of these variants for population 1 variance is σ_1^2 by n_1 for population 2 the variance is σ_2^2 by n_2 . So, the variance is if we want to know the difference of the variance of the 2 population we have to add the variance.

We know that the formula variance of $A - B$ equal to variance of A + variance of B . So, the formula if I say variance of $(A - B)$ you might have studied. So, variance of A plus variance plus variance of B . So, for the first population variances Sigma 1 square by n_1 for second population the variance is Sigma 2 square by n_2 so this is the variance of this population which is in blue in color.

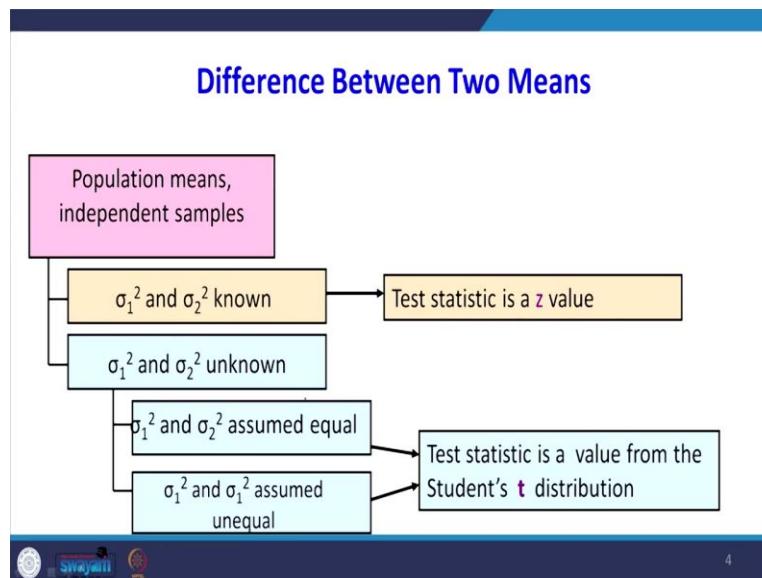
(Refer Slide Time: 03:14)



So, will you use this result in coming this one this is the classification of different 2 sample test. One classification is population means for independent samples, population mean for dependent samples, population proportions and population variances. In the population mean for independent variables will compare group 1 versus group 2 both the populations are independent. The population mean for dependent variables same group before and after the treatment so this is dependent samples.

In population proportion there are 2 population we will take proportion one from population 1 versus variance of proportion 2 we will take another population. Similarly for comparing variances of 2 population we will find the variance of 1 of population 1 versus variance of 2 of population 2.

(Refer Slide Time: 04:10)

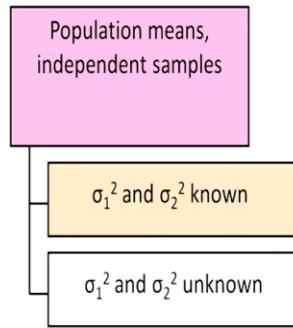


First you will start with between 2 means so the population means for independent sample there are 2 possibilities there one is the sigma1 square and sigma2 square is known sigma1 square is variance of population 1 Sigma 2 square is variance of population 2. The another category is variance are unknown for population 1 and 2 variants are unknown. When variant are unknown if we can assume it is equal or we can assume it it is not equal.

If variance of population 1 and 2 is known we should go for test statistics is a Z value if the variance is unknown we have to go for t statistics.

(Refer Slide Time: 04:57)

σ_1^2 and σ_2^2 Known



Assumptions:

- Samples are randomly and independently drawn
- both population distributions are normal
- Population variances are known

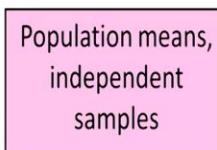


5

We will see what are the assumptions for when a variance of 2 populations are known to you, the first assumption is samples are randomly and independently drawn both the populations are normal. We have seen in the first slide both the populations are normal population variance are known but different.

(Refer Slide Time: 05:17)

σ_1^2 and σ_2^2 Known



When σ_1^2 and σ_2^2 are known and both populations are normal, the variance of $\bar{X}_1 - \bar{X}_2$ is

$$\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

...and the random variable

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has a standard normal distribution



6

Then sigma1 square and sigma2 square known we will find out what is the variance of that population when sigma1 square and sigma2 square are known both the populations are normal the variance of the difference in the variance of population 1 and 2 is nothing but the summation of their variance that is Sigma 1 square by n 1 plus Sigma 2 square by n 2 which I have already

explained. So, corresponding Z statistics is $(X \bar{1} - X \bar{2}) - (\mu_1 - \mu_2)$ divided by $\sqrt{(\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2})}$ has a standard normal distribution.

So, if it is a 1 sample what was our formula if it is a 1 sample our formula was like this Z equal to $(X \bar{1} - \mu) / \sigma / \sqrt{n}$ the denominator σ / \sqrt{n} is called standard error. So, in this formula previously for one sample you have taken only one sample here there is 2 sample that is we are finding the difference of the 2 sample so it should be $X \bar{1} - X \bar{2}$. Previously we have assumed one mean population mean now we are going to find the difference of the 2 population mean we are going to assume the difference of 2 population so see $\mu_1 - \mu_2$ this standard error σ / \sqrt{n} is we got from this one.

So, this you have to take the square root of this so that will become $\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}$, when the population means our independent samples first we will start from null hypothesis. So, now that $\mu_1 - \mu_2$ will have some difference D_0 , the test statistics for $(\mu_1 - \mu_2)$ is $(X \bar{1} - X \bar{2}) - D_0$ the previously what we had taken view we wrote $\mu_1 - \mu_2$ as it is but instead of some $\mu_1 - \mu_2$ we can write only the difference also here, so $\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}$.

(Refer Slide Time: 07:19)

Hypothesis Tests for Two Population Means

Two Population Means, Independent Samples

Lower-tail test: $H_0: \mu_1 \geq \mu_2$ $H_1: \mu_1 < \mu_2$ i.e., $H_0: \mu_1 - \mu_2 \geq 0$ $H_1: \mu_1 - \mu_2 < 0$	Upper-tail test: $H_0: \mu_1 \leq \mu_2$ $H_1: \mu_1 > \mu_2$ i.e., $H_0: \mu_1 - \mu_2 \leq 0$ $H_1: \mu_1 - \mu_2 > 0$	Two-tail test: $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$ i.e., $H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 \neq 0$
--	--	--

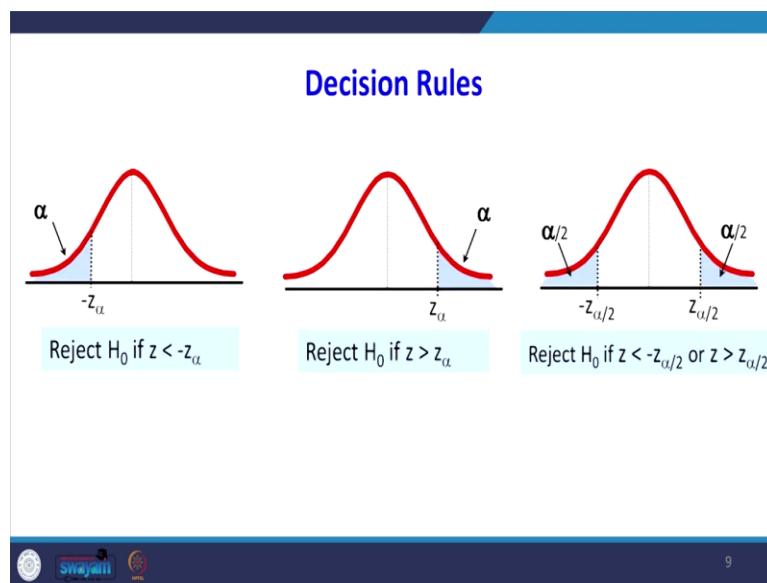
8

When both the Sigma are known to us there are different types of test as possible it may be a lower tailed test how we are saying lower tails is the left one you see that. The null hypothesis is

$\mu_1 \geq \mu_2$, alternate hypothesis is $\mu_1 < \mu_2$. So, this will be a left tailed test otherwise you can bring this μ_2 to the left side so it will be $\mu_1 - \mu_2 \geq 0$, alternate hypothesis is $\mu_1 - \mu_2 < 0$ that means. So, H_0 : is $\mu_1 - \mu_2$, H_1 : is $\mu_1 > \mu_2$, we can find the difference $\mu_1 - \mu_2$ less than or equal to 0, $\mu_1 - \mu_2$ greater than 0 it is a right tailed test.

Here we can have an assumption that μ_1 equal to μ_2 and μ_1 not equal to μ_2 , if you bring on left hands to be $\mu_1 - \mu_2$ equal to 0 then $\mu_1 - \mu_2$ not equal to 0. We will see this different test pictorially.

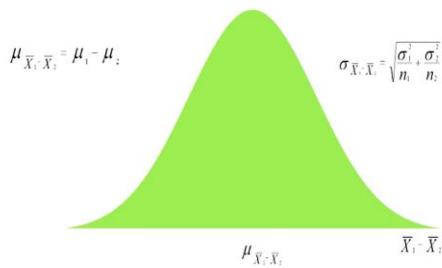
(Refer Slide Time: 08:28)



The decision rule for the left tailed test is reject H_0 if the calculated Z value is less than $-Z$ alpha. The decision rule for the right tailed test is reject H_0 if the calculated Z value is greater than Z alpha, for two tail test reject H_0 if Z values less than $-Z$ alpha by 2 are greater than Z alpha by 2, anyone can happen.

(Refer Slide Time: 08:58)

Hypothesis Testing about the Difference in Two Sample Means



10

So this distribution is sampling distribution of difference of 2 samples. So, the mean of this distribution is mu of ($\bar{X}_1 - \bar{X}_2$) bar the standard deviation of this distribution is Sigma 1 square by n 1 + Sigma 2 square by n 2 this already I have explained how we have got this values.

(Refer Slide Time: 09:23)

Sampling Distribution of $\bar{x}_1 - \bar{x}_2$

- Expected Value $E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$
- Standard Deviation (Standard Error) $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

where: σ_1 = standard deviation of population 1

σ_2 = standard deviation of population 2

n_1 = sample size from population 1

n_2 = sample size from population 2



11

For the sampling distribution of difference of 2 population mean the expected value is E of $\bar{X}_1 - \bar{X}_2$ equal to $\mu_1 - \mu_2$. So, the standard deviation is already we have seen root of Sigma 1 square by n 1 plus Sigma 2 square by n 2.

(Refer Slide Time: 09:43)

Interval Estimation of $\mu_1 - \mu_2$: σ_1 and σ_2 Known

- Interval Estimate

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where: $1 - \alpha$ is the confidence coefficient



12

The interval estimate is $X_1 - X_2 \pm Z_{\alpha/2}$ (root of ($\sigma_1^2/n_1 + \sigma_2^2/n_2$)) this was for 2 sample. If it is a one sample what was the remember that it is X bar plus or - $Z_{\alpha/2}$ σ by root n , so equation just as extended for 2 sample population. So, instead of X bar we are going to write $X_1 - X_2 \pm Z_{\alpha/2}$ will be as it is this σ by root n is replaced by $\sigma_1^2/n_1 + \sigma_2^2/n_2$ so what I am trying to say is that even though it is a 2 sample Z test the logic of extending one sample to 2 sample is very easy you need not remember the formula just intuitively you can extend the formula.

(Refer Slide Time: 10:52)

Problem (σ_1 and σ_2 Known)

- A product developer is interested in reducing the drying time of a primer paint.
- Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time.
- From experience, it is known that the standard deviation of drying time is 8 minutes, and this inherent variability should be unaffected by the addition of the new ingredient.
- Ten specimens are painted with formulation 1, and another 10 specimens are painted with formulation 2; the 20 specimens are painted in random order.
- The two-sample average drying times are $\bar{x}_1 = 121$ minutes and $\bar{x}_2 = 112$ minutes, respectively.
- What conclusions can the product developer draw about the effectiveness of the new ingredient, using $\alpha = 0.05$?

Source: Applied Probability and Statistics for Engineers by Douglas C. Montgomery and George C. Runger John Wiley, 3rd Ed. 2003



13

We will do problem on this when σ_1 and σ_2 is known the problem is taken from this book applied probability and statistics for engineers by Montgomery. A product developer is

interested in reducing the drying time of a primer paint 2 formulations of the painter tested. Formulation one is the standard chemistry and formulation 2 has new drying ingredient that should reduce the drying time. From experience it is known that the standard deviation of drying time is 8 minutes and this inherent variability should be unaffected by the addition of new ingredient.

10 specimens are painted with the formulation 1 and another 10 specimens are painted with the formulation 2. The 20 specimens are painted in random order to sample average drying times are \bar{X}_1 bar is 121 minutes \bar{X}_2 bar is 112 minutes respectively. What conclusions can the product developer draw about the effectiveness of the new ingredient?

(Refer Slide Time: 12:12)

Problem (σ_1 and σ_2 Known)

1. The quantity of interest is the difference in mean drying times, $\mu_1 - \mu_2$, and $\Delta_0 = 0$.
2. $H_0: \mu_1 - \mu_2 = 0$, or $H_0: \mu_1 = \mu_2$.
3. $H_1: \mu_1 > \mu_2$. We want to reject H_0 if the new ingredient reduces mean drying time.
4. $\alpha = 0.05$
5. The test statistic is

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where $\sigma_1^2 = \sigma_2^2 = (8)^2 = 64$ and $n_1 = n_2 = 10$.

14

Assuming significance level alpha equal to 0.05, the first step in the hypothesis testing is the quantity of interest is the difference in the mean of drying time so $\mu_1 - \mu_2$ if there is no difference it will become 0. Next we will form null hypothesis $\mu_1 - \mu_2 = 0$ or μ_1 equal to μ_2 that means the mean drying time of ingredient 1 and 2 is same. The alternative hypothesis is $\mu_1 > \mu_2$ what we are going to assume that the new ingredient is more efficient for that the drying time is going to be less.

If it is less will be greater and μ_2 will be lesser so we are writing μ_1 greater than μ_2 that is going to be our alternative hypothesis. We want to reject H_0 if the new ingredient reduces

mean drying time for alpha 5% test statistic is $X_1 - X_2$ bar - mu 1 - mu 2 that is 0, root of Sigma 1 square by n 1 + Sigma 2 square by n 2 standard deviations given 8 so the variance is 64 the sample size is n 1 n 2 equal to 10 when you supply this Sigma 1 square Sigma 2 square n 1 and n 2 in this formula.

(Refer Slide Time: 13:30)

Problem (σ_1 and σ_2 Known)

6. Reject $H_0: \mu_1 = \mu_2$ if $z_0 > 1.645 = z_{0.05}$.
7. Computations: Since $\bar{x}_1 = 121$ minutes and $\bar{x}_2 = 112$ minutes, the test statistic is

$$z_0 = \frac{121 - 112}{\sqrt{\frac{(8)^2}{10} + \frac{(8)^2}{10}}} = 2.52$$

15

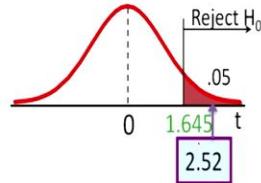
So the rejection rule is if mu1 equal to mu2 that is a null hypothesis, reject H0 if the test statistics is greater than 1.645, how we got this 1.645 because you see that this test is right tailed test so right tailed test means it will be like this. So, when alpha equal to 0.05 corresponding Z value is 1.645 if the calculated Z is greater than 1.645 we have to reject our null hypothesis. Next we will compute our calculate our Z average test statistics.

After supplying X1 and X2 we are getting 2.52, so 2.52 will be in the rejection side so we have reject we have to reject null hypothesis.

(Refer Slide Time: 14:29)

Problem (σ_1 and σ_2 Known)

$$t = \frac{(121 - 112) - 0}{\sqrt{8^2 \left(\frac{1}{10} + \frac{1}{10} \right)}} = 2.52$$



Decision:
Reject H_0 at $\alpha = 0.05$

Conclusion:
There is evidence of a difference in means.



16

You see that 2.52 is lying on the rejection site that decision is we are to reject null hypothesis by comparing the critical value.

(Refer Slide Time: 14:44)

Problem (σ_1 and σ_2 Known)

8. Conclusion: Since $z_0 = 2.52 > 1.645$, we reject $H_0: \mu_1 = \mu_2$ at the $\alpha = 0.05$ level and conclude that adding the new ingredient to the paint significantly reduces the drying time. Alternatively, we can find the P -value for this test as

$$P\text{-value} = 1 - \Phi(2.52) = 0.0059$$

Therefore, $H_0: \mu_1 = \mu_2$ would be rejected at any significance level $\alpha \geq 0.0059$.



17

The same problem can be done with help of comparing p values also so what conclusion we can have since the Z calculated is that is 2.52 is greater than 1.645 we reject H_0 that is μ_1 equal to μ_2 at the Alpha equal to 0.05 level and conclude that adding new ingredient to the paint significantly reduce the drying time. Since we reject null hypothesis we are going to accept ultra 2 hypothesis that says that new ingredient is reducing the drying time.

Alternatively we can find the p-value for this test, so because it is a right-tailed test the p-value should be 1 minus when calculated Z value is 2.5 minus corresponding probability so we will get we got 0.0059, I will verify this with the help of Python how we got this 0.0059, see these 0.0059 we are comparing with alpha. So, when we have to reject they see that therefore H₀ mu 1 equal to mu 2 would be rejected to any significance level alpha is greater than 0.0059 what is happening here the value of p is very less. So we are to reject our null hypothesis.

(Refer Slide Time: 15:59)

```

Problem ( $\sigma_1$  and  $\sigma_2$  Known)

In [2]: import pandas as pd
import numpy as np
import math
from scipy import stats

In [6]: def Z_and_p(x1,x2,sigma1,sigma2,n1,n2):
    z = (x1-x2)/(math.sqrt(((sigma1**2)/n1)+((sigma2**2)/n2)))
    if(z < 0):
        p = stats.norm.cdf(z)
    else:
        p = 1 - stats.norm.cdf(z)
    print (z,p)

In [7]: Z_and_p(121,112,8,8,18,10)
2.5155764746872635 0.00594189462107364

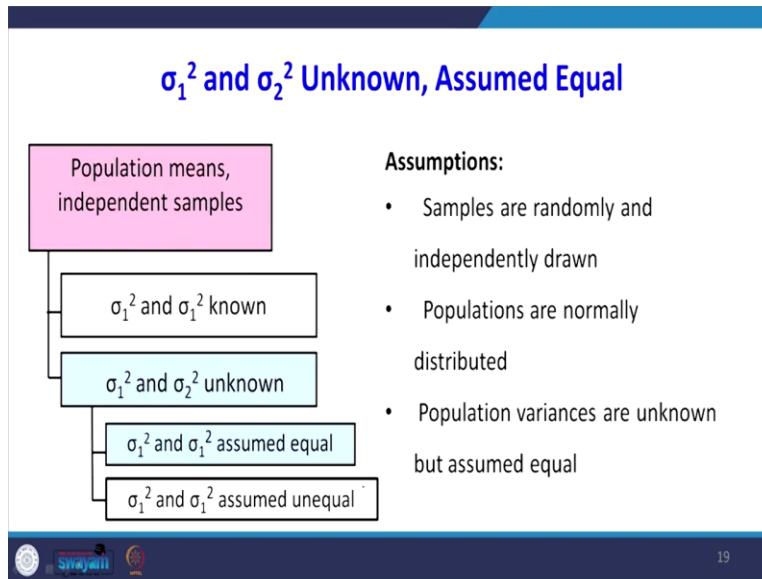
```

So we have done your Python code for this import pandas as pd import numpy np import math from scipy import stats we are going to make one definition here we are going to define your function define def that is a standard syntax Z _ and_ p the variable which are going to take is X 1 X 2 Sigma 1 Sigma 2 n 1 n 2 with useful notations. Then we will find out the Z value Z value is X 1 - X 2 square root of Sigma 1 square by n 1 + Sigma 2 square by n 2 the value of Z is less than 0 then p value is nothing but the actual value you can read as it is.

When the Z is it is like this if the Z value is coming on the left hand side the p value you can read as it is but if the if the z values positive we need the right side area so 1 minus the left side area will give you the right side area p equal to 1 – stats. Norm. cdf z, so print Z p. So, this code in our monitor just you can try this after pausing this video you can verify this answer. So, Z underscore Z underscore and underscore P just and supply all the values.

In the previous problem the X_1 bar is 121, X_2 bar is 112 Sigma 1 is the 8 Sigma 2 is 8 n_1 is 10 n_2 is 10, what happening we are getting the Z value is 2.51 there is a Z calculated value look at this p value and you go to previous slide say p value here also we got 0.0059, so here we get with the help of Python. When you compare with the alpha this is very small we are rejecting null hypothesis.

(Refer Slide Time: 18:01)



Now we will go to the second category of the problem. So, far in the previous problem we know Sigma 1 square, Sigma 2 square but this case Sigma 1 square, Sigma 2 square is unknown but we are assuming it is equal. There is a concept behind why we are assuming it is equal whenever we make comparison we can compare our the comparison is meaningful only when the variance of 2 groups are equal like comparing the performance of third year versus fourth year student is there is no meaning for that.

We can compare only the third year student versus another third year students so that way the variance has to be equal then only the comparison will be meaningful. So, the second case you see the blue 1 Sigma 1 square and Sigma 2 square unknown when done we are going to make another assumption that it is equal there is another possibility it is unknown but unequal we will come to the that one in the after sometime.

First we will go Sigma 1 square Sigma 2 square unknown but assumed equal. What are the assumption we are making samples are randomly and independently drawn populations are normally distributed population variances are unknown but assumed to be equal.

(Refer Slide Time: 19:15)

σ₁² and σ₂² Unknown, Assumed Equal

- The population variances are assumed equal, so use the two sample standard deviations and pool them to estimate σ
- use a t value with $(n_1 + n_2 - 2)$ degrees of freedom

20

The population variance are assumed equal use the 2 sample standard deviation and pool them to estimate the variance or standard deviation use your t value with $n_1 + n_2 - 2$ degrees of freedom. Actually what the concept here is assume that there is a group 1 this variance is S_1^2 square suppose n_1 sample there is a group 2 the variance is S_2^2 square this sample sizes n_2 . Suppose assuming population variance are equal you can pull the variance how we can do the pull the variance.

We can find out the weighted variance that we are going to called pooled variance the weighted variance is nothing but suppose assume that we are going to find out the weighted mean. So, what is the formula for weighted mean suppose $W_1 X_1 + W_2 X_2$ divided by sum of weighted $W_1 + W_2$ this is nothing but your weighted mean. Here the weight is nothing but your degrees of freedom. Suppose for the sample 1 the degrees of freedom is $n_1 - 1$ here the variance is S_1^2 square plus sample to the weighted is corresponding degrees of freedom S_2^2 square.

So, next we have to sum the degrees of freedom $n_1 - 1 + n_2 - 1$ that is nothing but $n_1 + n_2 - 2$ so, this is nothing but the pooled variance.

(Refer Slide Time: 20:59)

Test Statistic, σ_1^2 and σ_2^2 Unknown, Equal

The test statistic for

$\mu_1 - \mu_2$ is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

Where t has $(n_1 + n_2 - 2)$ d.f.,

$$\text{and } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

21

The test statistics for $\mu_1 - \mu_2$ is say previously we are used as Z now we are using t, $((\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)) / (\text{root of } (s_p^2/n_1 + s_p^2/n_2))$, since s_p^2 is same we can bring left hand side that is nothing but pooled variance that pooled variances see that $(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2$ divided by $(n_1 + n_2 - 2)$ degrees of freedom. You see that here the degrees of freedom is $n_1 + n_2 - 2$.

(Refer Slide Time: 21:38)

Decision Rules

Two Population Means, Independent Samples, Variances Unknown

Lower-tail test:

$$\begin{aligned} H_0: \mu_1 - \mu_2 &\geq 0 \\ H_1: \mu_1 - \mu_2 &< 0 \end{aligned}$$

Upper-tail test:

$$\begin{aligned} H_0: \mu_1 - \mu_2 &\leq 0 \\ H_1: \mu_1 - \mu_2 &> 0 \end{aligned}$$

Two-tail test:

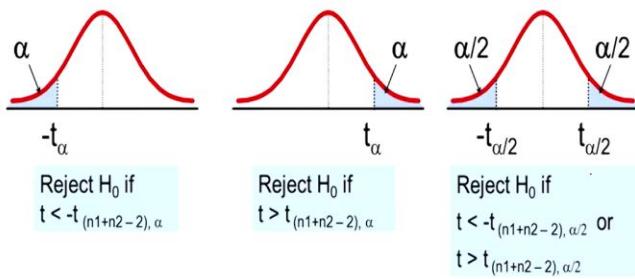
$$\begin{aligned} H_0: \mu_1 - \mu_2 &= 0 \\ H_1: \mu_1 - \mu_2 &\neq 0 \end{aligned}$$

22

Then we do the population mean standard deviation as unknown the error also there is a possibility to see this test is left tailed test this is right tailed test this is 2 tailed test.

(Refer Slide Time: 21:46)

Decision Rules



23

The next slides see that if the $-t_{\alpha}$ no α this one this is a left tailed test this is the right tail test middle one the right hand side was is the 2 tailed test.

(Refer Slide Time: 21:57)

σ_1^2 and σ_2^2 Unknown, Assumed equal

- Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process.
 - Specifically, catalyst 1 is currently in use, but catalyst 2 is acceptable.
 - Since catalyst 2 is cheaper, it should be adopted, providing it does not change the process yield.
 - A test is run in the pilot plant and results in the data shown in table.
 - Is there any difference between the mean yields?
 - Use 0.05, and assume equal variances.
- | Observation Number | Catalyst 1 | Catalyst 2 |
|--------------------|------------|------------|
| 1 | 91.50 | 89.19 |
| 2 | 94.18 | 90.95 |
| 3 | 92.18 | 90.46 |
| 4 | 95.39 | 93.21 |
| 5 | 91.79 | 97.19 |
| 6 | 89.07 | 97.04 |
| 7 | 94.72 | 91.07 |
| 8 | 89.21 | 92.75 |
- $\bar{x}_1 = 92.255$ $\bar{x}_2 = 92.733$
 $s_1 = 2.39$ $s_2 = 2.98$

24

We will take one problem where Sigma 1 square and Sigma 2 square unknown assumed equal, 2 catalyst are being analyzed to determine how they affect the mean yield of a chemical process. Specifically catalyst 1 is currently in use but catalyst 2 is acceptable. Since catalyst 2 is cheaper it should be adopted providing it does not change the process yield a test run in the pilot plant and the result in the data shown in the table.

Is there any difference between mean yields use alpha equal to 0.05 and assume equal variances. By looking at this problem you see that how it is given the variance are equal, no where the population variance is given so we should go for sample t-test assuming equal variance.

(Refer Slide Time: 22:58)

σ_1^2 and σ_2^2 Unknown, Assumed equal

1. The parameters of interest are μ_1 and μ_2 , the mean process yield using catalysts 1 and 2, respectively, and we want to know if $\mu_1 - \mu_2 = 0$.
2. $H_0: \mu_1 - \mu_2 = 0$, or $H_0: \mu_1 = \mu_2$
3. $H_1: \mu_1 \neq \mu_2$
4. $\alpha = 0.05$
5. The test statistic is

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

• 25

As usual the step one is we have to see which parameter of the population we are studying whether mean or variance. Now it is a mean so the parameter of interest are mu1 and mu2 the mean process yield using catalyst one and 2 respectively and we want to know if mu1 - mu2 equal to 0. So, $H_0: \mu_1 - \mu_2 = 0$, so $H_1: \mu_1 \neq \mu_2$, alpha 0.05 there is the test status is t_0 $\bar{X}_1 - \bar{X}_2$ minus the difference in the mean as you would mean s_p this is so s_p square was inside root we brought left side so it is the pooled standard deviation root of 1 by $n_1 + 1$ by 2.

(Refer Slide Time: 23:47)

σ_1^2 and σ_2^2 Unknown, Assumed equal

6. Reject H_0 if $t_0 > t_{0.025,14} = 2.145$ or if $t_0 < -t_{0.025,14} = -2.145$.
7. Computations: From Table 10-1 we have $\bar{x}_1 = 92.255$, $s_1 = 2.39$, $n_1 = 8$, $\bar{x}_2 = 92.733$, $s_2 = 2.98$, and $n_2 = 8$. Therefore

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(7)(2.39)^2 + 7(2.98)^2}{8 + 8 - 2} = 7.30$$
$$s_p = \sqrt{7.30} = 2.70$$

and

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{2.70\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{92.255 - 92.733}{2.70\sqrt{\frac{1}{8} + \frac{1}{8}}} = -0.35$$



26

So, what will happen when we look at the say statistical table when 14 degrees of freedom because it is a 2 tail, so since it is 2 tail but this area is 0.025 this area is 0.025 when 14 degrees of freedom the right hand side value is 2.145, I am writing here at the bottom it is 2.145 the left hand side it is -2.145 it is symmetric so positive or negative from the previous slide we have the mean of sample one is 92.255 and standard deviation of the sample one is 2.39 for n_1 equal to 8.

Similarly for the sample 2 the sample mean is 92.733 a standard deviation is 2.98 and n_2 is 8 therefore first we will find the pooled variance by using the formula $(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2$ divided by $n_1 + n_2 - 2$ after substituting this value we are getting 7.30 we could take the square root of that will get the standard deviation so use 2.70.

(Refer Slide Time: 25:10)

σ_1^2 and σ_2^2 Unknown, Assumed equal

8. Conclusions: Since $-2.145 < t_0 = -0.35 < 2.145$, the null hypothesis cannot be rejected. That is, at the 0.05 level of significance, we do not have strong evidence to conclude that catalyst 2 results in a mean yield that differs from the mean yield when catalyst 1 is used.



27

In this t formula we are getting -0.35 obviously you have to locate this is the rejection region. So, what we are concluding since -2.145 is less than that what we are got the value is going on the left hand side that is -0.35 the calculated t value is -0.35 so in this the -0.35 will be the acceptance side. So, we have to accept null hypothesis so what is happening -0.35 the null this cannot be rejected that is at the 5% level of significance we do not have strong evidence to conclude that the catalyst 2 result in a mean yield that differs from mean yield when catalyst 1 is used.

(Refer Slide Time: 25:59)

σ_1^2 and σ_2^2 Unknown, Assumed equal

```
In [12]: b =[ 89.19,90.95,90.46,93.21,97.19,97.04,91.07 , 92.75]
In [13]: a =[ 91.5, 94.18,92.18,95.39,91.79,89.07,94.72,89.21]
In [14]: stats.ttest_ind(a, b, equal_var = True)
Out[14]: Ttest_indResult(statistic=-0.3535908643461798, pvalue=0.7289136186068217)
In [21]: stats.t.ppf(0.025,14) #critical t value
Out[21]: -2.1447866879169277
```



28

So, now with the help of Python when Sigma 1 square, Sigma 2 square unknown assuming equal variance will solve the problem. Previously I am taking the b equal to I am assigning into an

objective b I have taken an array a, I have taken the next one. So, when he stats dot t-test underscore independent you call that array a, b equal variance and equal to true right it can be true or false the next after sometime we will solve that problem when it is a true directly we are getting you see that the test statistics - 0.35, so the p value is 0.72.

So we have to accept our null hypothesis we can see how we got - 2.144 also stats.t.ppf if you want to know the key value 0.025 when 40 degrees of freedom so we can compare t values also see that so when you see that one so the t value is - 2.14 but our test statistics - 0.35 it is lying on the acceptance region we are accepting our null hypothesis. Dear students so for what we have seen we are comparing hypothesis testing for 2 sample.

We have seen three types of problems number one is Sigma 1 square, Sigma 2 square is known then we have compared to the mean. Another type of problem is Sigma 1 square and Sigma 2 square is unknown and we have assumed equal variance. We have done the Z test and we have done the t-test and also I have explained the concept behind of standard deviation of the difference of 2 population variance.

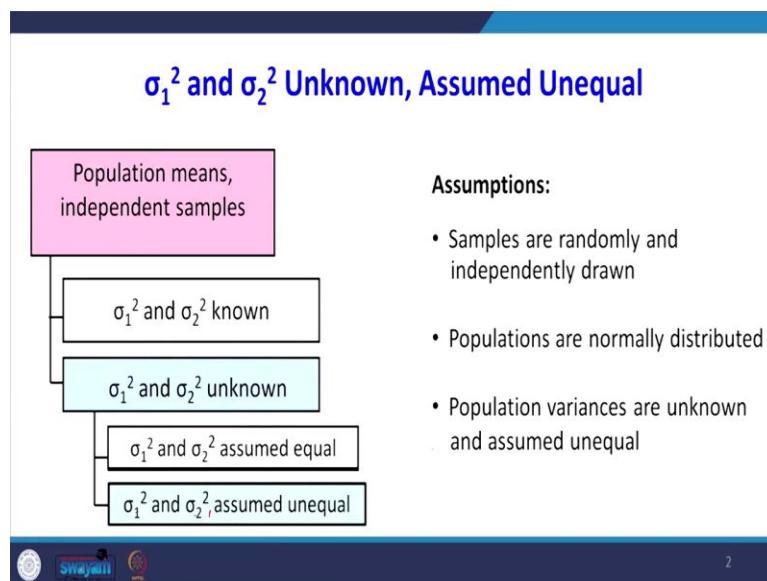
So what is the concept there if I want to know the difference of the 2 population variance you have to add the variance. If you want to know the difference of the 2 population mean just you can find the difference of the population mean. In the next class we will take a new problem where Sigma 1 square Sigma 2 square unknown but if it is unequal variance with that we will start the next class, thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 21
Hypothesis Testing: Two Sample Test-II

Dear students in the previous class we have seen the problem in comparing hypothesis testing in to population where σ_1^2 and σ_2^2 are known. Then we have seen that the next problem that is σ_1^2 and σ_2^2 is unknown but assumed equal variance. In this class we are going to take another category of the problem where σ_1^2 and σ_2^2 are unknown but assumed unequal.

(Refer Slide Time: 00:59)



What are the assumptions we are having the samples are randomly and independently drawn the populations are normally distributed population variance are unknown and assumed unequal. The population variances are assumed unequal so your pooled variance is not appropriate. So, use here we have to use a t-value with new deals of freedom the formula for degrees of freedom is this one nu(v) equal to $((s_1^2 / n_1) + (s_2^2 / n_2))^2$ divided by $((s_1^2 / n_1)^2 / (n_1 - 1)) + ((s_2^2 / n_2)^2 / (n_2 - 1))$

(Refer Slide Time: 01:30)

Test Statistic: σ_1^2 and σ_2^2 Unknown, Unequal

The test statistic for

σ_1^2 and σ_2^2 unknown

σ_1^2 and σ_2^2 assumed equal

σ_1^2 and σ_2^2 assumed unequal

$\mu_1 - \mu_2$ is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where t has v degrees of freedom:

$$v = \frac{\left[\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^{-1} \right]}{\left(\frac{s_1^2}{n_1} \right)^2 / (n_1 - 1) + \left(\frac{s_2^2}{n_2} \right)^2 / (n_2 - 1)}$$



4

The t statistics is $X_1 \text{ bar} - X_2 \text{ bar} - \text{the difference root of } ((s_1^2 / n_1) + (s_2^2 / n_2))$ you see the previous problem we have used S_P square by $n_1 + S_P$ square by n_2 where the variances are equal. But here it is unequal we cannot use S_P square in both the places so we have to use only s_1 square, the corresponding formula for degree of freedom already which we explained.

(Refer Slide Time: 01:57)

Problem: Test Statistic: σ_1^2 and σ_2^2 Unknown, Unequal

- Arsenic concentration in public drinking water supplies is a potential health risk.
- An article in the Arizona Republic (Sunday, May 27, 2001) reported drinking water arsenic concentrations in parts per billion (ppb) for 10 metropolitan Phoenix communities and 10 communities in rural Arizona.
- The data as shown:

	Metro Phoenix	Rural Arizona	
Phoenix,	3	Rimrock,	48
Chandler,	7	Goodyear,	44
Gilbert,	25	New River,	40
Glendale,	10	Apache Junction,	38
Mesa,	15	Buckeye,	33
Paradise Valley,	6	Nogales,	21
Peoria,	12	Black Canyon City,	20
Scottsdale,	25	Sedona,	12
Tempe,	15	Peyson,	1
Sun City,	7	Casa Grande,	18

Reference: Applied statistics and probability for engineers, Douglas C. Montgomery, George C. Runger, John Wiley & Sons, 2007

$$\bar{x}_1 = 12.5$$

$$s_1 = 7.63$$

$$\bar{x}_2 = 27.5$$

$$s_2 = 15.3$$



5

We will take you one sample problem will solve this one the problem is arsenic concentration in public drinking water supplies is a potential health risk. An article in Arizona Republic Sunday May 27 2001 reporter drinking water arsenic concentration in parts per billion ppb for 10 metropolitan phoenix communities and 10 communities in rural Arizona are given in the table. We can know what is the X_1 bar that is a 12.5, s_1 is 7.63, X_2 bar is 27.5, s_2 is 15.3.

(Refer Slide Time: 02:38)

Problem: Test Statistic: σ_1^2 and σ_2^2 Unknown, Unequal

- We wish to determine if there is any difference in mean arsenic concentrations between metropolitan Phoenix communities and communities in rural Arizona.

6

We wish to determine if there is any difference in mean arsenic concentration between metropolitan Phoenix communities and communities in rural Arizona.

(Refer Slide Time: 02:48)

Problem: Test Statistic: σ_1^2 and σ_2^2 Unknown, Unequal

1. The parameters of interest are the mean arsenic concentrations for the two geographic regions, say, μ_1 and μ_2 , and we are interested in determining whether $\mu_1 - \mu_2 = 0$.
2. $H_0: \mu_1 - \mu_2 = 0$, or $H_0: \mu_1 = \mu_2$
3. $H_1: \mu_1 \neq \mu_2$
4. $\alpha = 0.05$ (say)
5. The test statistic is

$$t_0^* = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

7

So what are the steps in hypothesis testing as usual the first step is the parameter of interest are the mean arsenic concentration for the 2 regions say μ_1 and μ_2 and we are interested in determining whether $\mu_1 - \mu_2$ equal to 0. So, what will be about null hypothesis null hypothesis is $\mu_1 - \mu_2$ equal to 0 otherwise μ_1 equal to μ_2 . Alternative hypothesis μ_1 not equal to μ_2 , because the signs are complementary so alpha is 5% but is not given we have

to assume it is a 5% is the formula for test statistics is this one t_0 is $X_1 - X_2$ bar root of $s_1^2/n_1 + s_2^2/n_2$

(Refer Slide Time: 03:40)

Problem: Test Statistic: σ_1^2 and σ_2^2 Unknown, Unequal

6. The degrees of freedom

$$v = \frac{\left[\left(\frac{s_1^2}{n_1} \right) + \left(\frac{s_2^2}{n_2} \right) \right]^2}{\left(\frac{s_1^2}{n_1} \right)^2 / (n_1 - 1) + \left(\frac{s_2^2}{n_2} \right)^2 / (n_2 - 1)} = \frac{\left[\left(\frac{7.63^2}{10} \right) + \left(\frac{15.3^2}{10} \right) \right]^2}{\left(\frac{7.63^2}{10} \right)^2 / (10 - 1) + \left(\frac{15.3^2}{10} \right)^2 / (10 - 1)} = 13.2 \approx 13$$

Therefore, using $\alpha = 0.05$, we would reject $H_0: \mu_1 = \mu_2$ if $t_0^* > t_{0.025, 13} = 2.160$ or if $t_0^* < -t_{0.025, 13} = -2.160$

8

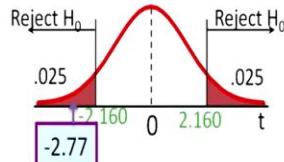
So, the degrees of freedom be all the data is given in the previous slide we can supply that value $s_1^2/n_1, s_2^2/n_2$ and so on. So, we are getting 13.2 approximately the degrees of freedom is 13 therefore using alpha 5% we would reject $H_0: \mu_1 = \mu_2$, if the p value the calculated T value is greater than 2.160 or p value is less minus because these values there are 2 ways we can get this value we can refer the T table but we can use Python also directly you can use the Python to get the critical value when alpha equal to 0.02 that means when the probability is 0.025 when degrees of freedom is 13.

(Refer Slide Time: 04:31)

Problem: Test Statistic: σ_1^2 and σ_2^2 Unknown, Unequal

7. Computations:

$$t = \frac{(12.5 - 27.5) - 0}{\sqrt{\left(\frac{7.63^2}{10} + \frac{15.3^2}{10}\right)}} = -2.77$$



Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is evidence of a difference in means.



9

So, we have done the t by using the calculated t values – 2.77 obviously – 2.77 is lying on the rejection side so we have to reject null hypothesis. So, what we are concluding there is evidence of difference in the means that means it is not the equal amount of arsenic is available there is it in some cities it is more in other cities it is less.

(Refer Slide Time: 05:00)

Problem: Test Statistic: σ_1^2 and σ_2^2 Unknown, Unequal

8. Conclusions: Because $t_0^* = -2.77 < t_{0.025, 13} = -2.160$,

- Reject the null hypothesis.
- There is evidence to conclude that mean arsenic concentration in the drinking water in rural Arizona is different from the mean arsenic concentration in metropolitan Phoenix drinking water.



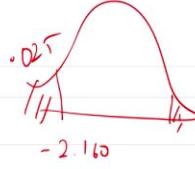
10

So, the conclusion because $t_0^* = -2.77$ is less than $t_{0.025, 13} = -2.160$ we have to reject a null hypothesis there is evidence to conclude that the mean arsenic concentration in the drinking water in rural Arizona is different from the arsenic concentration in metropolitan Phoenix drinking water it is not the same.

(Refer Slide Time: 05:22)

Problem: Test Statistic: σ_1^2 and σ_2^2 Unknown, Unequal

```
In [17]: stats.t.ppf(0.025,13) #critical t value  
Out[17]: -2.160368656461013  
  
In [18]: metro = [3,7,25,18,15,6,12,25,15,7]  
rural = [48,44,40,38,33,21,20,12,1,18]  
  
In [20]: stats.ttest_ind(metro,rural, equal_var = False)  
Out[20]: Ttest_IndResult(statistic=-2.7669395785560558, pvalue=0.015827284816100885)
```



11

So, we will use Python to solve this problem we can see the p value as I told you stats.t.ppf when in the t distribution when this area is 0.025 because a 2-tailed test area equal to 0.025 when the degrees of freedom is 13 we are getting it is - 2.160 so it is a - 2.160 our calculated t value is how much - 2.77 so - 2.77 will be on the left-hand side obviously we have to reject it. Instead of doing that it is very simple in Python you take array 1 as the values which is given for Metro there are another one array 2 that is call it is rural.

The value is given a rural area so stats.ttest_ind(call this to array metro, rural equal underscore variance you have to type equal to equal to false, do you remember for a previous one we have written is it true. Now simply write false you'll get the your t value your p value obviously it is a 2-tailed test, so, the p value the alpha is it is very small the p value is very small so we have to reject a null hypothesis when compared to alpha it is only 0.01 so we have to reject our null hypothesis.

(Refer Slide Time: 06:42)

Dependent Samples

Tests Means of 2 Related Populations

- Paired or matched samples
- Repeated measures (before/after)
- Use difference between paired values:

$$d_i = x_i - y_i$$

- Assumptions:
 - Both Populations Are Normally Distributed



12

Now we will go to another setup problem where there are samples are dependent. So, the test of 2 related populations they are called paired sample or match the samples so it is repeated measures. The same population we are collecting the data before and after so we have to find use the difference between paired sample $d_i = x_i - y_i$. So, what the logic is, is this is say population 1 this is population this is also population 1 same population before what was the this before any treatment suppose the we can see a lot of hair oil advertisements are coming before applying oil what was the length of your hair.

Here you can see some example some values who take some sample mean after sometime up after applying hair oil you can see what was the say this is X_1 bar this is Y_1 , X bar Y bar before and after we find when you plot the difference, so you take X_1 from the sample one before then you take Y_1 from the same sample because it is the independent sample when you plot the difference that when you keep on collect different pair from the same sample when you plot the difference that will follow normal distribution. So both our populations are normally distributed.

(Refer Slide Time: 08:14)

Test Statistic: Dependent Samples

The test statistic for the mean difference is a **t value**, with $n - 1$ degrees of freedom:

$$t = \frac{\bar{d} - D_0}{\frac{s_d}{\sqrt{n}}} \quad \bar{d} = \frac{\sum d_i}{n} = \bar{x} - \bar{y}$$

D_0 = hypothesized mean difference

s_d = sample standard dev. of differences

n = the sample size (number of pairs)



13

The test statistics for the mean difference is the t value with $n - 1$, degrees of freedom. So, here you see previously we would write X bar here the mean of the difference there is a d bar, you add all the difference divided by n there is nothing but equal to X bar - Y bar this was difference in the population sd is for that data for the difference the data what was the standard deviation the root of n. so, we will get the t-value.

(Refer Slide Time: 08:46)

Decision Rules: Dependent Samples

Lower-tail test:

$$H_0: \mu_1 - \mu_2 \geq 0$$

$$H_1: \mu_1 - \mu_2 < 0$$

Upper-tail test:

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

Two-tail test:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

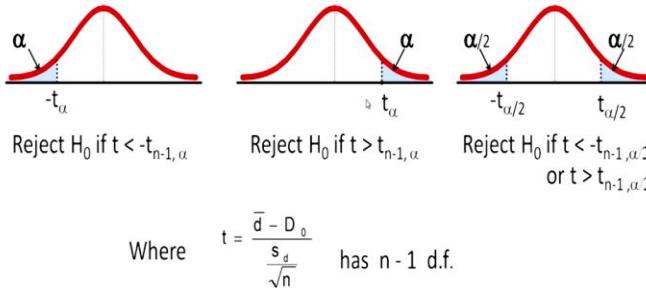


14

Here also see this is your left tailed test the second one is right tailed test this is 2-tailed test.

(Refer Slide Time: 08:51)

Decision Rules: Dependent Samples



15

Left tail right tail 2 tail test but only the t is $(\bar{d} - D_0) / (s_d / \sqrt{n})$ with $n - 1$ degrees of freedom.

(Refer Slide Time: 09:04)

Dependent Samples: Example

- An article in the Journal of Strain Analysis (1983, Vol. 18, No. 2) compares several methods for predicting the shear strength for steel plate girders.
- Data for two of these methods, the Karlsruhe and Lehigh procedures, when applied to nine specific girders, are shown in Table .
- We wish to determine whether there is any difference (on the average) between the two methods.

Reference: Applied statistics and probability for engineers, Douglas C. Montgomery, George C. Runger, John Wiley & Sons, 2007

16

We will take one example for dependent sample an article in the journal of strain analysis compares that is volume 18 and number 22 compare several methods of predicting the shear strength of steel plate girders. Data for 2 of these methods one method is Karlsruhe another method is Lehigh procedures when applied to nine specific graders are shown in the table. I think these these 2 methods are the different way of measuring the shear strength.

We wish to determine whether there is any difference on the average value between 2 methods because the populations are same 2 difference are conducted.

(Refer Slide Time: 09:49)

Table : Strength Predictions for Nine Steel Plate Girders (Predicted Load/Observed Load)			
Girder	Karlsruhe Method	Lehigh Method	Difference dj
S11	1.186	1.061	0.119
S21	1.151	0.992	0.159
S31	1.322	1.063	0.259
S41	1.339	1.062	0.277
S51	1.200	1.065	0.138
S21	1.402	1.178	0.224
S22	1.365	1.037	0.328
S23	1.537	1.086	0.451
S24	1.559	1.052	0.507

17

So, called Karlsruhe method Lehigh method in Karlsruhe method this was the values this is Lehigh method this was the value. You are finding the differences look at this here the difference are positive there is a possibility the difference may be negative also that will subtracted from the positive value there is no problem.

(Refer Slide Time: 10:09)

Inferences About the Difference Between Two Population Means: Matched Samples

1. The parameter of interest is the difference in mean shear strength between the two methods, say, $\mu_D = \mu_1 - \mu_2 = 0$.
2. $H_0: \mu_D = 0$
3. $H_1: \mu_D \neq 0$
4. $\alpha = 0.05$
5. The test statistic is

$$t_0 = \frac{\bar{d}}{s_D/\sqrt{n}}$$

18

So, the first step is the parameter of interest is the difference in the mean shear strength between 2 methods there's $\mu_D = \mu_1 - \mu_2$ equal to see rather we call it as difference is equal to 0 the third would third step is $\mu_D \neq 0$ so alpha equal to 5% those the tested statistics is d bar divided by (S_D / \sqrt{n}) it is nothing but the same thing previously what was the t formula if it is the if it is

not paired sample what was the t formula ($\bar{X} - \mu$) divided by (S/\sqrt{n}) but the \bar{X} bar is nothing but the mean of the difference.

This is nothing but the standard deviation of the difference this was the difference in the mean all others are same. Because what I am saying every statistical test has some link once that is why you have to follow the order of learning this statistics because if you in between if you are going for some lectures that may require certain prerequisites. So, when you learn this one so you have to follow this sequence so that will be very easy for connecting with other statistical test.

(Refer Slide Time: 11:28)

**Inferences About the Difference Between Two Population Means:
Matched Samples**

6. Reject H_0 if $t_0 > t_{0.025, 8} = 2.306$ or if $t_0 < -t_{0.025, 8} = -2.306$.

7. Computations: The sample average and standard deviation of the differences d_j are $\bar{d} = 0.2736$ and $s_D = 0.1356$, so the test statistic is

$$t_0 = \frac{\bar{d}}{s_D/\sqrt{n}} = \frac{0.2736}{0.1356/\sqrt{9}} = 6.05$$

19

When you look at the table when the Alpha value is 0.025 that is half of the Alpha values 0.025 80 degrees of freedom so that value is 2.3, so if the calculated 2 value is it is like this so what will you do this value on positive side we are getting 2.306 and negative side we are getting – 2.306 this is the value which you got from the table. The calculated t values lies on either side of this limit it will be rejected. so, what we got the mean of the difference is 0.2736 the standard deviation is 0.1356 when you input this data we are getting 6.05 that is far away.

So we got to reject the null hypothesis when you reject null hypothesis the μ_1 , what was null hypothesis that $\mu_1 - \mu_2$ equal to 0, so H_1 : is $\mu_1 - \mu_2$ not equal to 0 when you reject that there is a difference if it is an hair oil example yes there is a effect of hair oil that help you to grow the hair.

(Refer Slide Time: 12:42)

8. Conclusions: Since $t_0 = 6.05 > 2.306$, we conclude that the strength prediction methods yield different results. The P -value for $t_0 = 6.05$ is $P = 0.0002$.

20

So, we are rejecting so we conclude that this strength prediction method yield different result we look at the p-value because with the help of statistical table especially t table find if the p-value is very difficult, but we will use Python to see what is the p-value.

(Refer Slide Time: 12:58)

```
In [37]: KARL=[1.186,1.151,1.322,1.339,1.200,1.402,1.365,1.537,1.559]
LEH=[1.061,0.992,1.063,1.062,1.065,1.178,1.037,1.086,1.052]

In [38]: stats.ttest_rel(KARL,LEH)
Out[38]: Ttest_relResult(statistic=6.0819394375848255, pvalue=0.00029529546278604066)
```

21

So, you take call it as array 1 call second one Lehigh stats dot t-test you see that this is underscore rel so that means dependent sample so call the 2 variable you will get this is your t value this is a p value less than alpha. So, we have to reject the null hypothesis so that means there is a difference.

(Refer Slide Time: 13:31)

Sampling Distribution of $\bar{p}_1 - \bar{p}_2$

- Expected Value

$$E(\bar{p}_1 - \bar{p}_2) = p_1 - p_2$$

- Standard Deviation (Standard Error)

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

where: n1 = size of sample taken from population 1

n2 = size of sample taken from population 2



Next we will go to another problem that is the inferences about the differences between 2 population proportions. So, far we talk, population proportion means whenever there is a categorical variable is there so far we have measured the continuous variable about from the population. If there is a categorical variable obviously the count is taken care that is nothing but the population proportions. So, inferences about the difference between 2 population proportions here also we can estimate the population proportion $p_1 - p_2$ we will do the hypothesis test about the difference of $p_1 - p_2$.

So, what is the expected value before going to this expected value say this is population 1 this is population 2, I take some sample from population 1 I am finding p_1 p_2 p_3 and so on. I am taking some sample from this population to from population 2 there are different sample. If every time if it take p_1 minus that is sample which is taken from sample 1, population 1 and population 2 if I find this difference $p_1 - p_2$ every time of a finding $p_1 - p_2$, so that difference if we plot that that will follow a normal distribution.

The same logic there if you want to know the difference of the variance for example here what was the variance you remember there pq/n , Sigma square is pq/n Sigma 1 square Sigma 2 square is when you call it as $p_1 q_1/n_1$ here it is $p_2 q_2/n_2$ if you want to know the difference in the variance you to add the variance. So, what will happen $(p_1 q_1/n_1) + (p_2 q_2/n_2)$

n_2 by n_2), this is the variance if you want to know those standard deviation just to take square root of that, that is why we have got this one.

So, the expected value nothing but the mean value of \bar{p}_1 bar - \bar{p}_2 bar is it is $p_1 - p_2$ standard deviation is Sigma of \bar{p}_1 bar - \bar{p}_2 is root of p_1 into $1 - p_1$ by $n_1 + p_2$ into $1 - p_2$ by n_2 this p_1 p_2 you see that previously you have taken this is nothing but this sample proportion p_1 p_2 also sample proportion for population 2, n_1 is size of the sample taken from population 1 n_2 is size of the sample taken from population 2.

(Refer Slide Time: 16:15)

Sampling Distribution of $\bar{p}_1 - \bar{p}_2$

- If the sample sizes are large, the sampling distribution of $\bar{p}_1 - \bar{p}_2$ can be approximated by a normal probability distribution.
- The sample sizes are sufficiently large if all of these conditions are met:

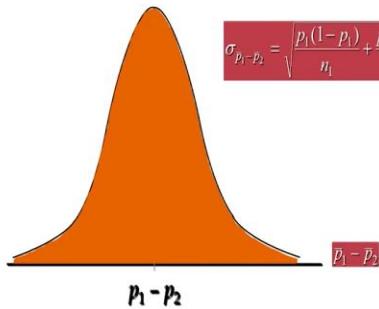
$n_1 p_1 \geq 5$ $n_1(1 - p_1) \geq 5$

$n_2 p_2 \geq 5$ $n_2(1 - p_2) \geq 5$

If the sample sizes are large the sampling distribution of \bar{p}_1 bar - \bar{p}_2 bar can be approximated by a normal probability distribution. The sample sizes are sufficiently large if all the conditions are met when np is greater than or equal to 5 or nq is greater than 5 then only we can approximate this one to the normal distribution.

(Refer Slide Time: 16:41)

Sampling Distribution of $\bar{p}_1 - \bar{p}_2$



You see that the mean of the mean of this non-word distribution is $p_1 - p_2$ the standard deviation is root of((p_1 into $1 - p_1$ by n_1) + (p_2 into $1 - p_2$ by n_2)).

(Refer Slide Time: 16:57)

Interval Estimation of $\bar{p}_1 - \bar{p}_2$

- Interval Estimation

$$\bar{p}_1 - \bar{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$$



The interval estimation is as usual \bar{p}_1 bar - \bar{p}_2 bar + or - $Z_{\alpha/2}$ root of \bar{p}_1 bar into $1 - \bar{p}_1$ by n_1 + \bar{p}_2 bar into $1 - \bar{p}_2$ by n_2 if it is a single sample you remember can you recollect what was the formula for interval estimation we evidently this way P bar + or - $Z_{\alpha/2}$ root of pq by n what is happening this pq is population proportion but we have to assume we have to approximate it with the sample proportion $p_1 q_1$ that is a small $p_1 q_1$ by n_1 . So what I am saying everything is came from work previous single sample hypothesis testing.

(Refer Slide Time: 17:41)

Point Estimator of the Difference Between Two Population Proportions

- p_1 = proportion of the population of households “aware” of the product after the new campaign
- p_2 = proportion of the population of households “aware” of the product before the new campaign
- \bar{p}_1 = sample proportion of households “aware” of the product after the new campaign
- \bar{p}_2 = sample proportion of households “aware” of the product before the new campaign

$$\bar{p}_1 - \bar{p}_2 = \frac{120}{250} - \frac{60}{150} = .48 - .40 = .08$$



We will take one problem point estimator of the difference between 2 population proportions say p_1 the proportion of population of households aware of the product after new campaign, p_2 is the proportion of population of households aware of the product before new campaign. So, we are going for new promotions we have to see the effectiveness of that promotion advertisement so \bar{p}_1 is the sample proportion of households aware of the product after the new campaign \bar{p}_2 is sample proportion of households aware of the product before the new campaigns.

So we will find the difference is any impact on the campaign new campaign on awareness. So, $\bar{p}_1 - \bar{p}_2$ is, we know that \bar{p}_1 is you know that this 120 divided by 250 because it is given so outer 250, 120 people are aware after the campaign, so, before the campaign out of 150 only 60 people are aware.

(Refer Slide Time: 18:55)

Hypothesis Tests about $p_1 - p_2$

- Hypothesis

We focus on tests involving no difference between the two population proportions (i.e. $p_1 = p_2$)

$H_0: p_1 - p_2 \geq 0$	$H_0: p_1 - p_2 \leq 0$	$H_0: p_1 - p_2 = 0$
$H_a: p_1 - p_2 < 0$	$H_a: p_1 - p_2 > 0$	$H_a: p_1 - p_2 \neq 0$
Left-tailed	Right-tailed	Two-tailed



So, the $p_1 - p_2$ bar is 8% so hypothesis we focus on test involving no difference between 2 population proportions. Here what is happening here they are also left tail test right tailed test 2 tail test even in the 2 sample population proportion also it can be left tail test right tailed test or 2 tail test.

(Refer Slide Time: 19:15)

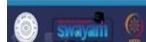
Hypothesis Tests about $p_1 - p_2$

- Standard Error of $\bar{p}_1 - \bar{p}_2$ when $p_1 = p_2 = p$

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

- Pooled Estimator of p when $p_1 = p_2 = p$

$$\bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2}$$



The standard error is root of p into $1 - p$ divided by n_1 1 by n_2 we have seen that one here also we can use pooled estimate of p when p_1 and p_2 equal to p . What is the meaning of this one is if the if you assume that the 2 population proportions are same then we can pool that so p bar is $n_1 p_1$ bar + $n_2 p_2$ bar it away $n_1 + n_2$.

(Refer Slide Time: 19:45)

Hypothesis Tests about $p_1 - p_2$

- Test Statistic

$$z = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$



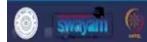
Test statistics is $(p_1 - p_2)/\sqrt{(p_1(1-p_1)/(n_1) + p_2(1-p_2)/(n_2))}$

(Refer Slide Time: 19:54)

Problem: Hypothesis Tests about $p_1 - p_2$

- Extracts of St. John's Wort are widely used to treat depression.
- An article in the April 18, 2001 issue of the *Journal of the American Medical Association* ("Effectiveness of St. John's Wort in Major Depression: A Randomized Controlled Trial") compared the efficacy of a standard extract of St. John's Wort with a placebo in 200 outpatients diagnosed with major depression.
- Patients were randomly assigned to two groups; one group received the St. John's Wort, and the other received the placebo.
- After eight weeks, 19 of the placebo-treated patients showed improvement, whereas 27 of those treated with St. John's Wort improved.
- Is there any reason to believe that St. John's Wort is effective in treating major depression? Use $\alpha = 0.05$.

Reference: Applied statistics and probability for engineers, Douglas C. Montgomery, George C. Runger, John Wiley & Sons, 2007



We will take on problem hypothesis test about $p_1 - p_2$ extract of st. John's Wort are widely used to treat depression this Jones what is a plant or medicine for treating depression. An article in April 18 2001 issue of Journal of American Medical Association the journal the article title is effectiveness of St. John's Wort major depression a randomized control trial. Compare the efficiency of standard extract of St. John's Wort with the placebo in 200 outpatients diagnosed with major depression. Patients were randomly assigned to groups one group received the st. John's Wort and other received the placebo.

After 8 weeks nine of the placebo treated patients showed input whereas 27 of those treated with St. John's Wort improved is there any reason to believe that St. John's Wort is effective curing major depression. Assume alpha equal to 5%. Now we have to see effect of this medicine and curing their depression.

(Refer Slide Time: 21:19)

Problem: Hypothesis Tests about $p_1 - p_2$

1. The parameters of interest are p_1 and p_2 , the proportion of patients who improve following treatment with St. John's Wort (p_1) or the placebo (p_2).

2. $H_0: p_1 = p_2$

3. $H_1: p_1 \neq p_2$

4. $\alpha = 0.05$

5. The test statistic is

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\hat{p}_1 = 27/100 = 0.27$, $\hat{p}_2 = 19/100 = 0.19$, $n_1 = n_2 = 100$, and

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{19 + 27}{100 + 100} = 0.23$$

 Swayam 

The parameter of interest are p_1 and p_2 the proportion of patients who improve following treatment with the st. John's what p_1 our placebo so the null hypothesis is there is no effect of this new medicine p_1 so we are going to assume $p_1 = p_2$ then alternative hypothesis it is not equal to p_2 okay, it does need not be 2-tailed test it is up to you to decide whether it is one tail or 2 tail test at present we are assuming we go that there is no difference in the medicine on the improvement of the patients.

The test statistics is $(p_1 \text{ hat} - p_2 \text{ hat}) / \sqrt{(\text{phat} (1 - \text{phat})) ((1/n_1) + (1/n_2))}$

so where p_1 is 27 by 100 p_2 19 by 100 n_1 n_2 is 100, so it is the pooled one so we see since the population proportions are same we are find out the pooled proportion $19 + 27 + 100 + 100$, 0.23.

(Refer Slide Time: 22:30)

Problem: Hypothesis Tests about $p_1 - p_2$

6. Reject $H_0: p_1 = p_2$ if $z_0 > z_{0.025} = 1.96$ or if $z_0 < -z_{0.025} = -1.96$.
7. Computations: The value of the test statistic is

$$z_0 = \frac{0.27 - 0.19}{\sqrt{0.23(0.77)\left(\frac{1}{100} + \frac{1}{100}\right)}} = 1.35$$

8. Conclusions: Since $z_0 = 1.35$ does not exceed $z_{0.025}$, we cannot reject the null hypothesis.
The P -value is $P \geq 0.177$. There is insufficient evidence to support the claim that St. John's Wort is effective in treating major depression.



34

So we have to reject our null hypothesis if it is greater than $+ 1.96$ otherwise less than $- 1.96$. so, the z value when you substitute it is 1.35 so what is happening 1.96 is here so this is the rejection region our 1.35 is lying on the acceptance region. So, what we are concluding since $Z_0 = 1.35$ does not exceed $Z_{0.025}$ that is 1.96 we cannot reject hypothesis. When you look at the p -value it is 0.177 so 0.177 is it is more than 0.5.

So we have to accept the null hypothesis there is insufficient evidence to support the claim that the Saint John's Wort is effective in treating major depression. So, we would accept our null hypothesis that means there is no evidence that Saint John's Wort is effective.

(Refer Slide Time: 23:44)

```
In [29]: import math
def two_samp_proportion(p1,p2,n1,n2):
    p_pool = ((p1*n2)+(p2*n1))/(n1+n2)
    x = (p_pool*(1- p_pool)*((1/n1)+(1/n2)))
    s = math.sqrt(x)
    z = (p1- p2)/s
    if (z < 0):
        p_val = stats.norm.cdf(z)
    else:
        p_val = 1 - stats.norm.cdf(z)
    return z, p_val**2
```

```
In [30]: two_samp_proportion(0.27,0.19,100,100)
Out[30]: (1.3442056254198995, 0.17888190308175567)
```

```
In [27]: stats.norm.cdf(1.3442056254198995)
Out[27]: 0.9105590484591222
```



35

This will do with the help of you can type in Jupiter this command then you have to verify this import math will make a function two_sample_proportions(p1, p2, n1, n2) first we will find out the pooled proportion with the help of Python we learn how to use 2 sample proportion test. So, import math we define a new function to underscore sample underscore proportion p 1 p 2 n1 n2 first we will find out the pool to proportion by $n_1 p_1 + n_2 p_2$ divided by $n_1 + n_2$ will solve with the help of Python 2 sample proportions hypothesis testing.

We know what is the formula? It is $p_1 - p_2$ divided by root of $p q$ into 1 by $n_1 + 1$ by n_2 . First there p we call it as a p underscore pool is nothing but pooled proportion where $p = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$ they will find the variance, variance is $p q$ multiplied by 1 by $n_1 + 1$ by n_2 then you will take a square root of that that will be the denominator of this formula that is a square root of X.

The value of $Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{pq(n_1 + 1)(n_2 + 1)}}$ since the value of capital $P_1 - P_2$ whatever assumed to 0 showing there so the Z value is $P_1 - P_2$ divided by your standard error. If the Z value is less than 0 so if the Z value is less than 0 we can find out the left side probability that is nothing but our p value. If it is greater than 0 we were to substrate from 1 so you will get the p value.

So in the given problem the p_1 population proportion is 0.27 p_2 population proportion is 1/9 so n_1 is 100 n_2 is 100. So, after getting we are getting the Z varies 1.3 the p value is 0.17 that is more than our 0.5, so we would accept our null hypothesis. Since stats, suppose if you want to know what was the Z critical value, so `stats.norm.cdf(1.35)` where we got this 1.35 so the corresponding probability 0.91 from this side this side is 0.91 will use Python to solve a 2 sample proportion test.

(Refer Slide Time: 26:59)

The screenshot shows a Jupyter Notebook interface with the title "Untitled1". The code cell contains the following Python script:

```

In [1]: import pandas as pd
import numpy
import math
from scipy import stats

In [2]: def Two_samp_proportion(p1,p2,n1,n2):
    p_pool = ((p1*n1)+(p2*n2))/(n1+n2)
    s_sq = (p_pool*(1-p_pool)*((1/n1)+(1/n2)))
    z = (p1-p2)/s
    if (z>0):
        p_val = stats.norm.cdf(z)
    else:
        p_val = 1 - stats.norm.cdf(z)

    return z, p_val*2

In [3]: Two_samp_proportion(0.27, 0.19, 100, 100)

```

We import pandas as pd import numpy import math from scipy we will import stats, so we will make your function, function name is to underscore samp underscore proportions p 1 p 2 n 1 n 2 first we will find the sample pooled the proportion by using this formula $n_1 p_1 + n_2 p_2$ divided by $n_1 + n_2$ then find out the variance is says p into $1 - p$ multiplied by $(1/n_1) + (1/n_2)$ so that will be the variance to get the standard deviation otherwise standard error will take square root of our variance that is s_{sq} .

So Z is $p_1 - p_2$, Z if the Z value is less than 0 the p value from the table we can treat as it is if the p value is positive we have to substrate from 1. So, when you the way we are going to call this function is by so the function will returns it p value that has to be multiplied by 2 because it is a 2-sample t-tests. So, we run this to sample proportion p 1 is 0.27 p 2 is 0.19 n 1 is hundred n 2 100 we will run it.

So we got the t value is 1.33 for the p value 0.17, so it is more than our alpha value, so we are to accept null hypothesis. Where this will conclude this will summarize this class we have seen 2 sample hypothesis testing when Sigma 1 square Sigma 2 square is unknown but not equal. Then we have seen 2 sample Z proportion test we have taken some problems then we solved it the next class will go for comparing 2 population variance using F test.

(Refer Slide Time: 00:37)

Agenda

- Comparing two population variances
- Choosing z or t test
- Sample size



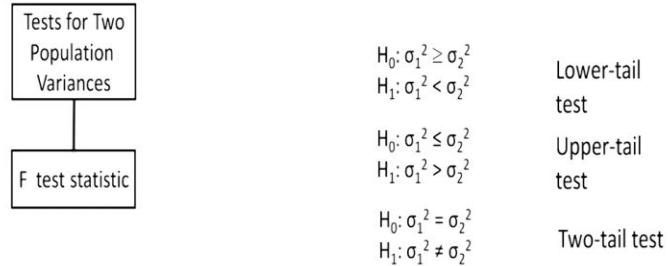
2

Welcome students in this class we will continue with the 2 sample hypothesis testing. In this class we will see how to compare population variance of 2 population so agenda for this class is comparing 2 population variances then many times student may have this doubt when to go for z test went to go for t test I will clarify that when to go for z test. The third one is it is the most important that what should be the sample size for doing any statistical analysis.

(Refer Slide Time: 00:58)

Hypothesis Tests for Two Variances

Goal: Test hypotheses about two population variances



The two populations are assumed to be independent and normally distributed

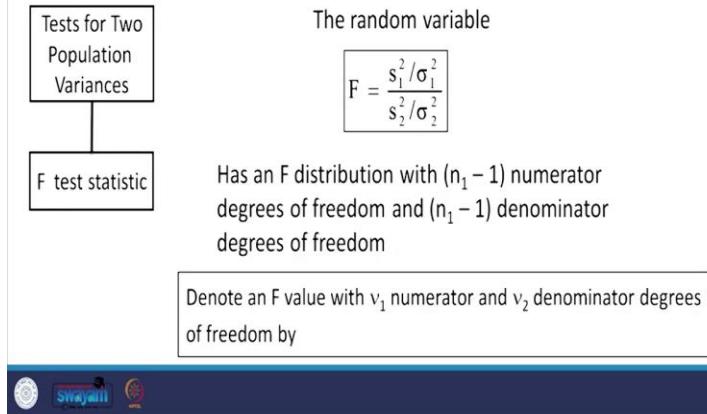
3

Now we will test the hypothesis test for 2 variances the goal is to hypothesis about population variances you see that the $H_0: \sigma_2^2 \leq \sigma_1^2$, the $H_1: \sigma_2^2 > \sigma_1^2$ it is over left tailed test otherwise the lower tail test it may be this way right it will be right skewed one what will happen this is the here also there is a left side this is left side test this is right side test this is 2 tailed test.

You have to remember that I did not draw the normal distribution this is a right skewed distribution this distribution is called F distribution. So, in the F distribution we have to find out the F statistics that F statistics will decide will help us to accept or reject null hypothesis. As usual here also there may be a left-tail test right tailed test or 2-tailed test but very important assumption which are which here to remember that the 2 populations are assumed to be independent and normally distributed.

(Refer Slide Time: 02:15)

Hypothesis Tests for Two Variances

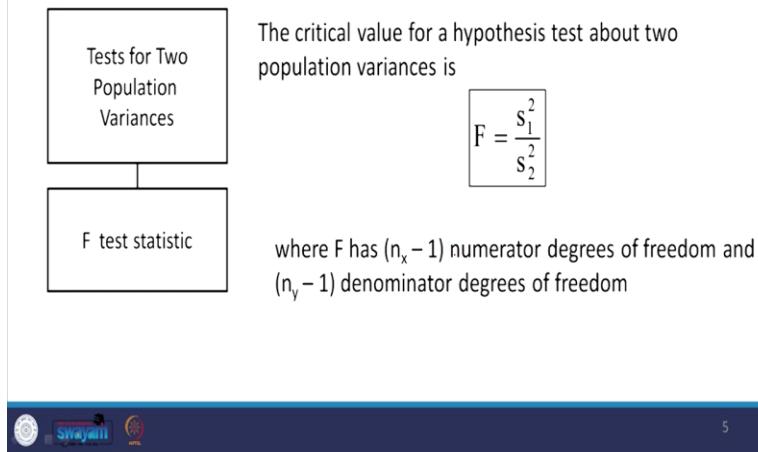


4

The test statistics for comparing 2 population variances F is nothing but $(s_1^2 / \sigma_1^2) / (s_2^2 / \sigma_2^2)$, the F statistic here is $(s_1^2 / \sigma_1^2) / (s_2^2 / \sigma_2^2)$ whole square. If you are assuming both the populations having equal variance the F will become s_1^2 / s_2^2 so it the F, $n_1 - 1$ numerator degrees of freedom and $n_2 - 1$ this is the $n_2 - 1$ denominator degrees of freedom.

(Refer Slide Time: 02:55)

Test Statistic

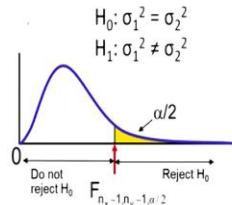
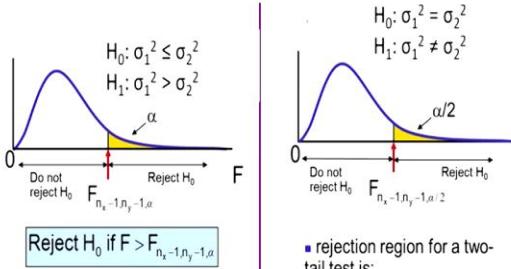


5

Yes I told you the critical value for a hypothesis test about 2 population variances you F equal to s_1^2 / s_2^2 what we are assuming here both population have equal variance where F was $n_1 - 1$ numerator degrees of freedom and this is $n_2 - 1$ denominator degrees of freedom.

(Refer Slide Time: 03:17)

Decision Rules: Two Variances



■ rejection region for a two-tail test is:

Reject H_0 if $F > F_{n_x-1, n_y-1, \alpha/2}$

where s_e^2 is the larger of the two sample variances



6

The decision rule for 2 variances for example most of the time the F test is only a right tailed test because whenever there is a variance so only we are bothered about only the upper limit of the variance the lower limit you will not bother about. Because taking lower limit is there is no meaning, so the what we have to assume that what is more important for is it should not exceed the upper limit of the variance.

It is like you see the bus has to come 9 o'clock if it has come 9 :05 or 9 :10 you will get bored but if it is coming early that there would not be any problem like that we have to bother about only the upper limit if there is a lower limit is not that much important. So, the degrees of freedom here is this is $n_1 - 1$ this is n_2 this is n_1 this is n_2 and other things for comparing 2 tailed test right as I told you for comparing 2 tailed test you see that this is σ_1^2 equal to σ_2^2 , square σ_1^2 not equal to σ_2^2 the rejection region for your 2 tailed test is see that this is $n_1 - 1, n_2 - 1$.

So what we have to do while finding the value of F right this is 1 while finding the value of F we have to maintain that the higher variance should be in the numerator. So, what we have to assume this s_1^2 is greater than s_2^2 so where s_1^2 is the larger of 2 sample variance that should go to in the numerator. If you take larger of 2 variants in the numerator you need not bother about the lower limit of the 2 tail test only we have to compare only upper limit of 2 tailed test for at accepting or rejecting null hypothesis.

(Refer Slide Time: 05:12)

Problem

- A company manufactures impellers for use in jet-turbine engines.
- One of the operations involves grinding a particular surface finish on a titanium alloy component.
- Two different grinding processes can be used, and both processes can produce parts at identical mean surface roughness.
- The manufacturing engineer would like to select the process having the least variability in surface roughness.
- A random sample of $n_1 = 11$ parts from the first process results in a sample standard deviation $s_1 = 5.1$ micro inches, and a random sample of $n_2 = 16$ parts from the second process results in a sample standard deviation of $s_2 = 4.7$ micro inches.
- We will find a 90% confidence interval on the ratio of the two standard deviations.

7

We will take one problem a company manufactures simpler for use in jet turbine engines one of the operations involves grinding a particular surface finish on a titanium alloy component 2 different grinding processes can be used and both processes can produce parts at identical means surface roughness. The manufacturing engineers would like to select the process having the least variability that is a point least variability in the surface roughness.

When you say generally the surface references measured by this way surface roughness is measured by this way suppose the surface roughness is this one so it is not good the surface roughness it cannot be perfectly smooth covered there should be a smaller variations. So, for the manufactures are also interested would like to select a process having least variability in the surface roughness a random sample of $n_1 = 11$ parts from first 2 processes result a sample standard deviation of 5.1 micro inches and random sample of into 16 parts from the second processes result in a sample standard deviation of $s_2 = 4.0$ micro inches.

We will find here 90% confidence interval on the ratio of 2 standard deviation then we will compare whether it is these variances are equal or not equal.

(Refer Slide Time: 06:42)

Problem

- Form the hypothesis test:
 $H_0: \sigma_1^2 = \sigma_2^2$ (there is no difference between variances)
 $H_1: \sigma_1^2 \neq \sigma_2^2$ (there is a difference between variances)
- Find the F critical values for $\alpha = .10/2$:

Degrees of Freedom:

- Numerator $F_{n_1-1, n_2-1, \alpha/2}$
(NYSE has the larger standard deviation):
 - $n_1 - 1 = 11 - 1 = 10$ d.f.
- Denominator:
 - $n_2 - 1 = 16 - 1 = 15$ d.f.



8

As usual first we have to form the null hypothesis the null hypothesis is $\sigma_2^2 = \sigma_1^2$ what is the meaning of this both the process are having equal variances. Alternative hypothesis is $\sigma_2^2 \neq \sigma_1^2$ there is there is a difference between variances so find the critical value alpha equal to 10%.

(Refer Slide Time: 07:06)

Problem

- Form the hypothesis test:
 $H_0: \sigma_1^2 = \sigma_2^2$ (there is no difference between variances)
 $H_1: \sigma_1^2 \neq \sigma_2^2$ (there is a difference between variances)
- Find the F critical values for $\alpha = .10/2$:

Degrees of Freedom:

- Numerator
 - $n_1 - 1 = 11 - 1 = 10$ d.f.
- Denominator:
 - $n_2 - 1 = 16 - 1 = 15$ d.f.



8

So the first thing is you have to find out the numerator degrees of freedom that is $n_1 - 1$ so $11 - 1$ 10 is the numerator degrees of freedom it is the denominator degrees of freedom $n_2 - 1$ so $16 - 1$ 15 is denominated degrees of freedom.

(Refer Slide Time: 07:24)

Problem

- Assuming that the two processes are independent and that surface roughness is normally distributed

$$\frac{s_1^2}{s_2^2} f_{0.95, 15, 10} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2} f_{0.05, 15, 10}$$

$$\frac{(5.1)^2}{(4.7)^2} 0.39 \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{(5.1)^2}{(4.7)^2} 2.85$$

or upon completing the implied calculations and taking square roots,

$$0.678 \leq \frac{\sigma_1}{\sigma_2} \leq 1.887$$



9

Assuming that the 2 process are independent and the surface reference is normally distributed we are going to find out the confidence interval of the ratio of their variance. So, what is the logic is suppose Sigma 1 square equal to Sigma 2 square, so it will in that interval 1 will be captured so F distribution is like this so here I am going to find out the confidence interval if you look at the F table the area is given only from right to left so when area equal to 0.05 because we have told it is a 10% see if the right side is 0.05 left side is 0.05.

If you look at the F table we can read the for 0.025 significance level what is the corresponding F value right if you want to know the left side 0.05 so you have to read 0.95 significance level then only you can find out the lower limit of the F. So what we see we write $(\sigma_1^2 / \sigma_2^2) \leq (s_1^2 / s_2^2)$, first time finding the upper limit upper limit is when it is a 0.05 see that when F equal to 0.05 numerated degrees of freedom is 15 denominated degrees of freedom is 10 that will be my upper limit.

The larger value of the variance should go to numerator. The left side the area the left side critical value is f 0.95 right if we want to because the F table read right to left 0.95 15, 10 degrees of freedom okay so s1 square 5.1, 5.1 divided by 4.7 then 0.95, 10 , 15 deals of freedom this value we are to read from the table. First we will read the upper limit so when the degrees of freedom is 0.05 numerator degrees of freedom is 15 and the denominator degrees of freedom is 10 we will see what is the F value.

(Refer Slide Time: 09:48)

Table of F-statistics P=0.05																
		t-statistics														
		F-statistics with other P-values: P=0.01 P=0.001														
		Chi-square statistics														
df2	df1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70	
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86	
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62	
6	5.99	5.14	4.76	4.53	4.39	4.23	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22	
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85	
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	

10

So numerator degrees of freedom is 15 denominator degrees of freedom is 10 see that this value is 2.85 that is where F value now what we are to do we have to know that lower limit for that when significance level is 0.95 when 15, 10 degrees of freedom we have to find out what is the F value for that purpose what you are to do we are to reverse the degrees of freedom so 10, 15 here 10, 15 is this point 2.54 because DF 1 do column says the numerator degrees of freedom the rows is denominated liters of freedom.

So this is 10, 15 so this 1 is 15, 10 because the column says the numerator this one is 10, 15 degrees of freedom when alpha equal to 0.05. So, if you want to know lower limit of their F so what you have to do?

(Refer Slide Time: 11:07)

Problem

- $f_{0.95,15,10} = 1 / f_{0.05,10,15} = 1/2.54 = 0.39$
- Since this confidence interval includes unity, we cannot claim that the standard deviations of surface roughness for the two processes are different at the 90% level of confidence.



11

You see that if you want to know the 0.95 for 15, 10 so we had to find the first we had to reverse the degrees of freedom 10, 15 then we had to find out alpha equal to 0.05 then we had to find the inverse of that so that is 1 divided by 2.54 how we got 2.54 this value 2.54 so 2.4 we got 0.39 so and going back again so that is why we got this 0.39 when you simplify this the lower limit is 0.678 the upper limit is 1.887 actually after taking the square root because upon completing the implied calculation take a square root of that we are getting this one.

You see that the range is capturing one that means there is a possibility Sigma 1 equal to Sigma 2 because there is this ratio implies that there is a possibility it may take one also so we have to accept our null hypothesis that is a Sigma 1 equal to Sigma 2. Since this conference interval includes unity we cannot claim that the standard deviation of surface roughness for the 2 process are different at the 90% level of confidence.

Going back in case the unity is not coming here so we cannot say both variants are equal since we are able to capture unity because lower limit is 0.6 upper limit is 1.8 there is a possibility the value of the ratio of Sigma 1 by Sigma 2 will become 1 so Sigma 1 by Sigma equal to 1 so that there is a possibility Sigma 1 will become equal to Sigma 2.

(Refer Slide Time: 12:57)

In [1]:

```
import pandas as pd
import numpy as np
import math
from scipy import stats
import scipy
```

In [45]:

```
scipy.stats.f.ppf(q=1-0.05, dfn= 15, dfd=10)
```

Out[45]:

```
2.8450165269958436
```

In [44]:

```
scipy.stats.f.ppf(q=0.05, dfn=15, dfd=10)
```

Out[44]:

```
0.3931252536255495
```

12

Now we will use a Python for finding the F table so for that you have to import pandas as pd import numpy as np, import math, from scipy import stats okay you can import scipy so what you have to do `scipy.stats.f.ppf` you have to give the probability so $1 - 0.05$ means 0.95, so 0.95 numerator degrees of freedom is 15 and 10 we are getting 2.84 so that value is nothing but 10, 15 2.84 `scipy.stats.f.ppf`, q equal to say $1 - 0.95$ numerator `dfn` is nothing but numerator degrees of freedom 15 denominator degrees of freedom 10 we are getting 2.84.

So, if you want to know the lower limit here we can directly read from the F table `scipy.stats.f.ppf` $q = 0.05$ numerator degrees of freedom is 15 denominator degrees of freedom is 10 you are getting 0.39 so this was our lower limit this was the our upper limit which previous problem.

(Refer Slide Time: 14:10)

F Test example:

```
In [9]: X = [3,7,25,10,15,6,12,25,15,7]
Y = [48,44,40,38,33,21,20,12,1,18]
import numpy as np

In [11]: F = np.var(X) /np.var(Y)
dfn = len(X) -1
dfd = len(Y) -1

In [12]: p_value = scipy.stats.f.cdf(F, dfn, dfd)

In [13]: p_value
Out[13]: 0.024680183438910465
```



13

So, for what you have done there are 2 group of population the standard deviation that is the variance of that 2 populations are given instead of that there is a possibility there is a population 1, X will be given Y will be given you have to find out the p value for that. Suppose I am assuming this is the population 1, I am going to call it as capital X there is another population 2, I am going to call it is capital Y, so my null hypothesis H_0 : I am going to assume $\sigma_x^2 = \sigma_y^2$. Alternative hypothesis is $\sigma_x^2 \neq \sigma_y^2$ right I am going to take alpha equal to 5 % that is 0.05.

So, declare variable X declare variable Y, import numpy as np then find out the ratio that is variance of X divided by variance of Y then you find numerator degrees of freedom len function that will tell you how many element is there in that array 1, so number of element - 1 that is a degrees of freedom for numerator, number of element - 1 there is a degrees of freedom for denominator. So, to get the p value equal to `scipy.stats.f.cdf` this is the syntax.

First declare what is F we have found this here then say what is it degrees of freedom numerator degrees of freedom denominator when you enter p value we are getting 0.024 so what will happen this your F distribution okay this is 0.05 this 0.024 see that is 0.024 is lesser than 0.05 so we have to reject a null hypothesis. When you reject it we are accepting that Sigma X square not equal to Sigma Y square. This is easiest way for testing variance of 2 population.

(Refer Slide Time: 16:27)

Z Vs t		
	σ -known	σ -unknown
$n \leq 30$	Z-test	t-test
$n > 30$	Z-test	Z-test Use Sample standard deviation

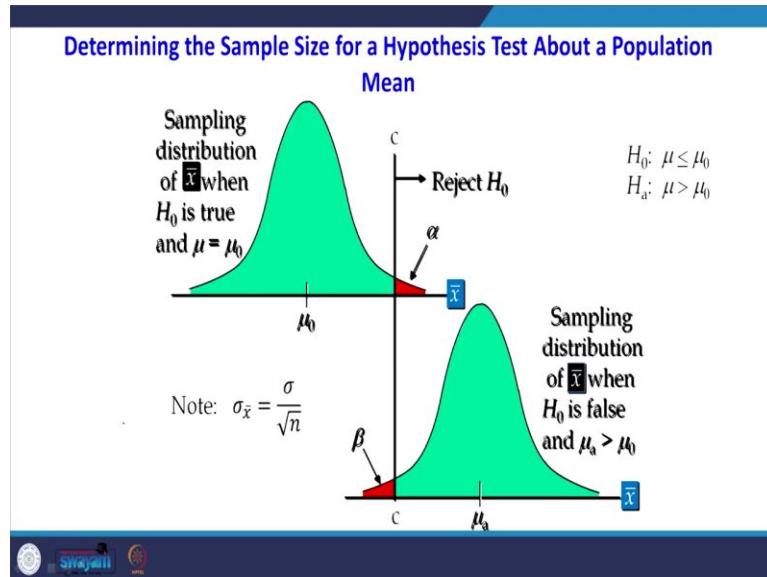
Many students may have doubt when we will go for Z test when you go t tests before going this I will summarize what you have done we have done one sample Z-test we have done t-test then we have done Z proportion test then we have done 2 sample z test then we have done 2 sample t-test. In the 2 sample t-test we have heard assumption population variance are equal are not equal. Then we have done 2 sample Z proportion test then later we have compared to sample variance test comparing we have compared to population variance.

After completing this then people may have doubt when should we go for Z test a t-test. You see that whenever the Sigma is known what is the meaning of the Sigma is known as whenever the population standard deviation is known you should go for Z test. So, to decide when should go for Z test there are 2 criteria one is the sample size and other is whether the Sigma is known or unknown. So, without considering the sample size whenever the Sigma is known whether the sample size is less than 30 or greater than 30 you should go for Z test.

So, when you look go for t test whenever Sigma is unknown and n is less than 30 you should go for t test. There may be a possibility Sigma is unknown but n is greater than 30 so instead of t test you can go for Z test because as we have studied previously the t distribution is the special case of your Z distribution. Whenever the degrees of freedom for example t distribution will be flat whenever the degrees of freedom is increasing, is increasing it will behave like your Z distribution.

So, whenever the Sigma is unknown n is greater than 30 you can go for we Z test that is why in many statistical package there would not separate tab for running Z test there will be a tab only for t test.

(Refer Slide Time: 19:01)



The another important question students will have is how to choose the sample size determining the sample size for a hypothesis the test about the population mean you see that there are 2 population mu₀ is the base population there is a mu_a is alternative mean, so this is a right tailed test where is a right tailed test if the mu greater than mu₀ you will reject it. Now what is happening any point which goes right hand side we will reject it.

This much portions which are falling acceptance side of my distribution I have accepted. Now this is beta there is a false acceptance this is alpha in correct rejection. By considering the Alpha and Beta you see in this point the value of X bar is same X bar for this population X bar for this population same.

(Refer Slide Time: 20:00)

Determining the Sample Size for a Hypothesis Test About a Population Mean

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_0 - \mu_a)^2}$$

where

z_α = z value providing an area of α in the tail

z_β = z value providing an area of β in the tail

σ = population standard deviation

μ_0 = value of the population mean in H_0

μ_a = value of the population mean used for the Type II error

Note: In a two-tailed hypothesis test, use $z_{\alpha/2}$ not z_α



So, by equating this X bar for both populations we can derive a formula for knowing the sample size considering Alpha and Beta this we can derive it, it is very simple derivation. So, what how to derive this we have to X bar because this line is same for both the population. So, here what will happen Z alpha is nothing but $(X$ bar - $\mu_0)$ divided by σ by \sqrt{n} for here Z beta equal to $(X$ bar - μ_a) divided by σ by \sqrt{n} .

You see that the value of this Z beta will become negative because this is lower value minus upper value, so when you equate this when you from these 2 from these 2 equation 1 and 2 when you simplify you can σ by \sqrt{n} we can get the value of n . So, when you the value of n is $(Z$ alpha + Z beta) 2 multiplied by $\sigma_2^2 \neq \sigma_1^2$ by $(\mu_0 - \mu_a)$ whole square okay because it is a 2 tailed test if it is a 2 tailed test we have to use Z alpha by 2 not Z alpha.

(Refer Slide Time: 21:24)

Determining the Sample Size for a Hypothesis Test About a Population Mean

- Let's assume that the manufacturing company makes the following statements about the allowable probabilities for the Type I and Type II errors:
- If the mean diameter is $\mu = 12$ mm, I am willing to risk an $\alpha = .05$ probability of rejecting H_0 .
- If the mean diameter is 0.75 mm over the specification ($\mu = 12.75$), I am willing to risk a $\beta = .10$ probability of not rejecting H_0 .



We will do a small problem for this let us assume that the manufacturing company makes the following statement about the allowable probability for type 1 type 2 error. Suppose somebody is manufacturing a shaft whose diameter say 50 mm. If the mean diameter is 50 mm I am willing to risk an alpha equal to 5% of a probability for rejecting null hypothesis if the mean diameter is 0.75 mm over the specification that is if the mu equal to 12.75.

I am willing to take a risk beta equal to 10% probability of not rejecting there is a false acceptance. Now what is happening alpha is given beta is given, alpha is your type 1 error, beta is a type 2 error, see actual mean is given an alternative mean is given.

(Refer Slide Time: 22:27)

Determining the Sample Size for a Hypothesis Test About a Population Mean

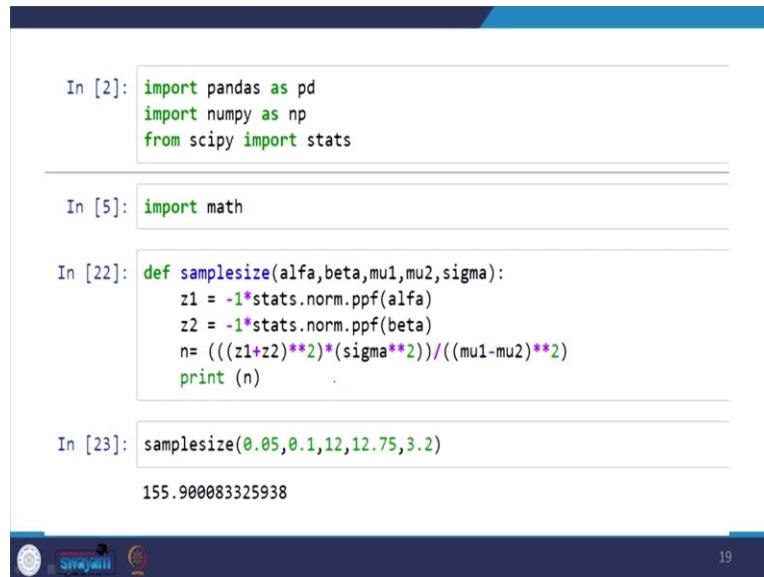
$$\begin{aligned}\alpha &= .05, \beta = .10 \\ z_\alpha &= 1.645, z_\beta = 1.28 \\ \mu_0 &= 12, \mu_a = 12.75 \\ \sigma &= 3.2\end{aligned}$$

$$\begin{aligned}n &= \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_0 - \mu_a)^2} = \frac{(1.645 + 1.28)^2 (3.2)^2}{(12 - 12.75)^2} \\ &= 155.75 \approx 156\end{aligned}$$



So, if that is the case what should be the sample size so alpha equal to 5% beta equal to 10% when alpha equal to 5% we can get it is 1.645 the Z_{beta} we can find out directly how to find out the Z_{beta} I am going back this value Z_{beta} is $X_{\bar{\mu}} - \mu$ divided by σ/\sqrt{n} all will be given. So, σ_{Beta} is 1.28, μ_0 is 12 μ_a is 12.75 σ is 3.2 just substitute this value so it should be 156.

(Refer Slide Time: 23:07)



```
In [2]: import pandas as pd
import numpy as np
from scipy import stats

In [5]: import math

In [22]: def samplesize(alfa,beta,mu1,mu2,sigma):
           z1 = -1*stats.norm.ppf(alfa)
           z2 = -1*stats.norm.ppf(beta)
           n= (((z1+z2)**2)*(sigma**2))/((mu1-mu2)**2)
           print (n)

In [23]: samplesize(0.05,0.1,12,12.75,3.2)
155.900083325938
```

We will use a Python for solving this problem import pandas pd import numpy as np from scipy import stats import math so we have to define a function samplesize (alpha beta mu 1 mu 2) Sigma so $Z_1 = -1 * \text{stats.norm.ppf}(\alpha)$, $Z_2 = -1 * \text{stats.norm.ppf}(\beta)$, $n = \frac{((Z_1 + Z_2)^2) * (\sigma^2)}{(\mu_1 - \mu_2)^2}$ that is the same formula if you substitute it you can print the n. So, when you supply alpha value beta value mu 1 value mu 2 value Sigma value we are getting 155.90.

So, far we what we have seen we have compared to 2 population then we have compared variance of the 2 population we have tested whether the populations are variants are equal or not then we have seen when to go for is a testament to when to go for t test after that by considering the Alpha and Beta value we have found your formula how to decide or how to arrive the value of your sample size it would take an old sample problem they have conducted then we found what was the value of the sample size, thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 23
ANOVA- I

Welcome students today we are, we will continue with the sample size calculation. After that we are going to see very important topic that is analysis of variance.

(Refer Slide Time: 00:37)

Agenda

- Sample Size Calculation
- One Way ANOVA – Introduction

2

Today's lecture planet sample size Calculation and one way Anova.

(Refer Slide Time: 00:41)

Determining Sample Size when Estimating μ

- Z formula

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$\frac{Z\sigma}{\sqrt{n}} = \bar{X} - \mu$$

$$\frac{Z\sigma}{(\bar{X} - \mu)} = \sqrt{n}$$

- Error of Estimation (tolerable error)

$$E = \bar{X} - \mu$$

- Estimated Sample Size

$$n = \frac{Z_a^2 \sigma^2}{E^2} = \left(\frac{Z_a \sigma}{E} \right)^2$$

$$\frac{Z\sigma}{(\bar{X} - \mu)} = n$$

- Estimated σ

$$\sigma \approx \frac{1}{4} \text{range}$$



Determining sample size with estimating Mu: We know that the Z formula is $(\bar{X} - \mu)$ divided by (σ/\sqrt{n}) . In this Sigma by root n, from this relationship this n is nothing but your sample size when you re-adjust this Mu, re-adjust this from this equations like $(Z\sigma)/\sqrt{n}$, $(\bar{X} - \mu)$, right then \sqrt{n} , can say $Z\sigma$ divided by $(\bar{X} - \mu)$, that will be your root n. You square both sides.

So, $(Z\sigma)^2$ divided by $(\bar{X} - \mu)^2$, that will be your n. That is nothing but this one. So, the numerator this $\bar{X} - \mu$ here we are going to call it as Error of estimation otherwise tolerable error. So, sample size is mu such that Z square, Sigma square, this is sigma square, Sigma square divided by error square ok. Since there, everywhere square is there we can bring it to common.

Many times, the value of population standard deviation may not be known. So there is approximation, one fourth of the range of the data which were collected can be taken as standard deviation.

(Refer Slide Time: 02:16)

Example: Sample Size when Estimating μ

$$E = 1, \sigma = 4$$

90% confidence $\Rightarrow Z = 1.645$

$$n = \frac{Z^2 \sigma^2}{E^2}$$
$$= \frac{(1.645)^2 (4)^2}{1^2}$$
$$= 43.30 \text{ or } 44$$



We will do an example, calculating the sample size. So, the permissible error is given. 1. The population standard deviation is 4. You want to conduct 90% confidence level. If it is 90% confidence level, this Z value is 1.645. Then you substitute here. You will get 43.30 that is nothing equivalent to 44.

(Refer Slide Time: 02:42)

Example

$$E = 2, \text{ range} = 25$$

95% confidence $\Rightarrow Z = 1.96$

estimated σ : $\frac{1}{4} \text{range} = \left(\frac{1}{4}\right)(25) = 6.25$

$$n = \frac{Z^2 \sigma^2}{E^2}$$
$$= \frac{(1.96)^2 (6.25)^2}{2^2}$$
$$= 37.52 \text{ or } 38$$



We will do another problem to find out the sample size. So, permissible error is 2. Range is given 25. 95% confidence level Z is equal to 1.96, this value which you have to get from the table. So, estimated Sigma is one fourth of the range. So it is 6.25 then substitute Z Square, Sigma square divided by E square we are getting 38.

(Refer Slide Time: 03:08)

Determining Sample Size when Estimating P

- Z formula

$$Z = \frac{\hat{P} - P}{\sqrt{\frac{P \cdot Q}{n}}}$$

- Error of Estimation (tolerable error)

$$E = \hat{P} - P$$

- Estimated Sample Size

$$n = \frac{Z^2 PQ}{E^2}$$

$$\begin{aligned} Z \sqrt{\frac{PQ}{n}} &= \hat{P} - P \\ Z^2 \frac{PQ}{n} &= (\hat{P} - P)^2 \\ \frac{Z^2 PQ}{n} &= E^2 \\ n &= \frac{Z^2 PQ}{E^2} \end{aligned}$$



Now, we will see how to find the sample size when estimating the population proportion. If you are estimating the population proportion, the formula for Z is different. \hat{P} - Capital P root of capital P and capital Q by n . This \hat{P} is the sample proportion capital P is the population proportion so the error is P - capital P . The same way what we have done previously. If you for example, when you bring Z root of PQ divided by n is equal to \hat{P} minus P . Square both sides.

Z square PQ by n equal to \hat{P} minus P , whole square. So, this will become Z Square PQ divided by \hat{P} minus P whole square. That is nothing but your n . So, \hat{P} - P we call it as E Square. So, n equal to Z Square PQ divided by E square.

(Refer Slide Time: 04:24)

Example

$$E = 0.03$$

$$98\% \text{ Confidence} \Rightarrow Z = 2.33$$

$$\text{estimated } P = 0.40$$

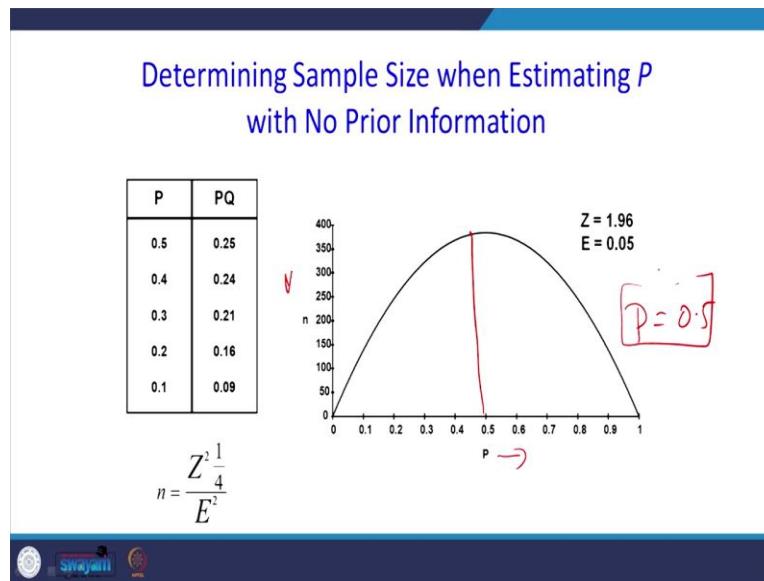
$$Q = 1 - P = 0.60$$

$$\begin{aligned} n &= \frac{Z^2 PQ}{E^2} \\ &= \frac{(2.33)^2 (0.40)(0.60)}{(0.03)^2} \\ &= 1,447.7 \text{ or } 1,448 \end{aligned}$$



We will do one problem here. So the permissible error is given. For 98% confidence level the Z value is 2.33 which you have to get it from the table. Estimated P is given population P is given 0.40. So Q is 0.60. You substitute this value it will be 1448 you see that whenever you go for population testing population proportion generally you ask yes or no type. That time obviously you have to have go for more number of samples. Sometimes what will happen the value of P which were given as 0.4 may not be known to you.

(Refer Slide Time: 05:08)



That time, how to decide the sample size, determining the sample size when estimating P with no prior information. Just look at this. The P is in the x-axis sample size in the y-axis suppose, the value of P equal to 0.5, the maximum sample size is required. So what the logic what we have to understand from here is if you are not knowing the population proportion we have to assume P equal to 0.5. Even the value of P is goes above 0.5 see that the, the n values decreasing.

The value of P is goes below 0.5 that time also the value of p is decreasing. It is better to assume P equal to 0.5 so that you will get maximum number of sample size.

(Refer Slide Time: 06:02)

Example

$$E = 0.05$$

90% Confidence $\Rightarrow Z = 1.645$

with no prior estimate of P , use $P = 0.50$

$$Q = 1 - P = 0.50$$

$$\begin{aligned} n &= \frac{Z^2 P Q}{E^2} \\ &= \frac{(1.645)^2 (0.50)(0.50)}{(0.05)^2} \\ &= 270.6 \text{ or } 271 \quad \checkmark \end{aligned}$$



In this situation, Error is given 90% is confidence level Z equal to 1.645 with no prior estimate of P we have to use $P = 0.50$. If we substitute P as 0.50 it will go for 271. The population proportion is not known to you or to take P value equal to 0.5.

(Refer Slide Time: 06:30)

Why ANOVA?

- We could compare the means, one by one using t-tests for difference of means.
- Problem: each test contains type I error
- The total type I error is $1 - (1 - \alpha)^k$ where k is the number of means.
- For example, if there are 5 means and you use $\alpha=.05$, you must make 10 two by two comparisons.
- Thus, the type I error is $1 - (.95)^{10}$, which is .4012.
- That is, 40% of the time you will reject the null hypothesis of equal means in favor of the alternative!



Next will go to next topic is Anova. Why Anova is required? So far we have seen two samples Z Test, two sample T test. Whenever there is a requirement for comparing more than two population. So far, what you have done? First, we have done one population next we have compared two population. If you want to compare more than two population, we can go for Anova. There is a possibility. You can compare 1 and 2 Suppose this is one, this is 2, this is 3, we can compare 1 and 2, 2 and 3, 1 and 3.

1 and 3 there are 3 comparison, but we will not go for that one because there is a reason for that. I will tell you we could compare the means one by one using T test of difference of means because what will happen each test contain Type 1 error. So the total type of error $1 - \alpha$ power k where K is number of means. For example, if there are 5 means if you use α equal to 0.05 because $5C_2$. You must, you must make 10 2 by 2 comparison because every comparison, your confidence level is 0.95.

If you are making 10 comparison the overall confidence level is 0.95 power 10. So $1 - \alpha$ confidence level is nothing but error. So, 0.95 power 10 that you substitute for 10 that is nothing but your error. That is 40 % that is, 40 percentage of the time you will reject the null hypothesis of equal means in favour of alternative. That is why we should not go for two samples T test whenever there if you want to compare more than two populations. We should go for Anova.

(Refer Slide Time: 08:32)

Hypothesis Testing With Categorical Data



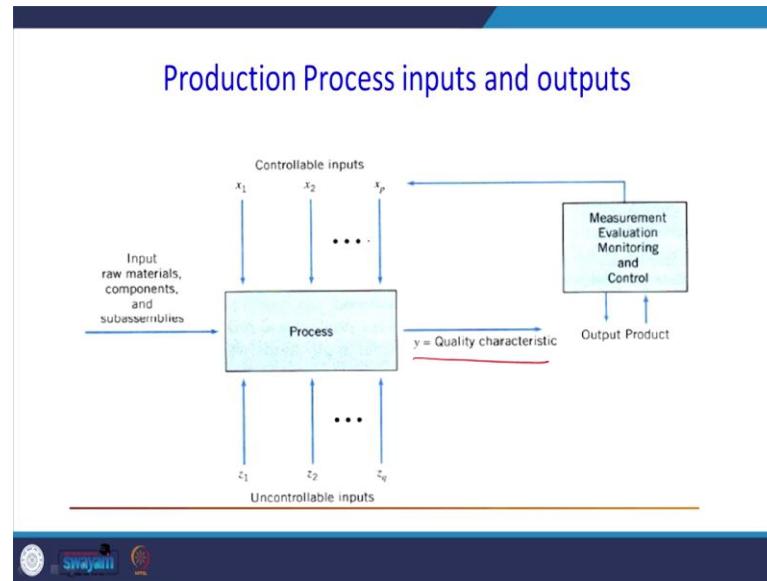
- Chi Square tests can be viewed as a generalization of Z tests of proportions
- Analysis of Variance (ANOVA) can be viewed as a generalization of t-tests: a comparison of differences of means across more than 2 groups.
- Like Chi Square, if there are only two groups, the two analyses will produce identical results – thus a t-test or ANOVA can be used with 2 groups



Hypothesis testing with categorical data because we are going to say Group 1, group2, group3. You see we have seen 2 samples Z proportion test. If you want to go for Z proportion test, if you want to go for comparing three population proportion, if you want to compare the proportion of more than two population, we should go for chi square test. Similarly, if you want to compare more than three population mean we should go for Anova.

So, chi-square test can be viewed as the generalization of Z test of proportion. The same way the anova can be viewed as the generalization of T test. The comparison of difference of mean across more than two groups like Chi square if there are only two groups, 2 analysis will produce identical result does a t test and Anova can be used with two groups. There are 2 groups we can go for Anova and a T test. Both will give you the same answer.

(Refer Slide Time: 09:38)



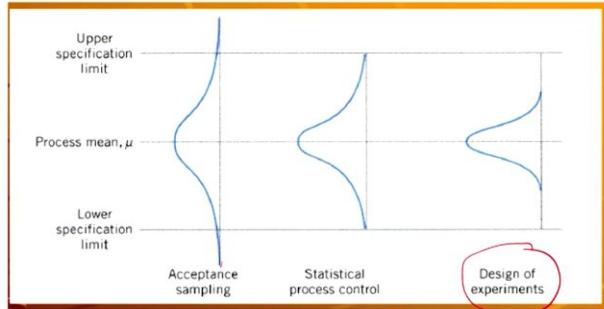
Why this concept of Anova is more important. Suppose, in the production process, there are different input variables. X_1, X_2 and X_P . These are controllable inputs. There are uncontrollable inputs, that is, Z_1, Z_2, Z_3 . So, the input maybe raw materials, component and subassemblies, output is the quality characteristics. What will happen this quality characteristics, Y is generally affected by this $X_1 X_2 X_3$ right. See that is output product.

If it is affecting Y we need to find out what combination of X_1 and X_2 will provide better Y value measurement evaluation monitoring and controlling for that purpose Anova is required. So the purpose of Anova is there are many input variables how this input variable is affecting the quality characteristics. For that we can find out with the help of Anova.

(Refer Slide Time: 10:42)

Application of quality-engineering techniques and the systematic reduction of process variability

n, c (10, 2)



13

See, when we want to improve the quality Application of quality engineering techniques and systematic reduction of process variability, because we want to improve the quality, is nothing but reduction of the variance. To go for acceptance sampling there will be lot of variance, because the acceptance sampling is a suppose, we go for say called n, c . If they say 10, 2 this is n is the lot size. C is acceptance number.

What is the meaning of this in n, c is the lot size 10, I will count all the defective products. The number of defective ways 2 or less, I will accept the whole lot. If the number of defective is two or more, I will reject the whole lot. So, there is a, it is a kind of intuitive process there is mathematics behind it which follow by normal distribution. But it is very easy way but there is more variability still will be maintained.

When we go for statistical process control, when the process started then we are controlling the process parameters that should go for statistical process control like our control chart. But the design of experiment days before starting of the experiment, before manufacturing in at laboratory level, you can see what are the parameters that will affect the quality of the product. So, we can control the product so that you can control the variable so that you can improve the quality.

With help of design of experiments, we can maintain high level of quality. That is why people are interested in design of experiments. The base for this design of experiment is nothing but your anova, analysis of variance. That is why I am trying to connect the connection between the design of experiments and Anova.

(Refer Slide Time: 12:40)

Effect of Teaching Methodology		
Group 1 Black Board	Group 2 Case Presentation	Group 3 PPT
4	2	2
3	4	1
2	6	3

I am going to explain the concept of analysis of variance with the help of an example. The example is, there are three teaching methodology. The one way of teaching is with help of blackboard and other ways with help of case studies. Third way is PowerPoint presentations suppose. I want to know which teaching methodology is more effective or is there any difference? Is there any influence of teaching methodology on student performance?

Totally 9 student was taken in each group 3 students are allotted randomly. So what do you see here is the marks obtained by the students when they are studying blackboard group. They ask to study case presentation group in where the teacher is using only PowerPoint presentations. Whatever value which you are saying, what value you are seeing this is the marks. These are the marks. What is the null hypothesis here?

(Refer Slide Time: 13:43)

$\bar{x}_1 = \frac{4+3+2}{3} = 3$
 $\bar{x}_2 = \frac{2+4+6}{3} = 4$
 $\bar{x}_3 = \frac{2+1+3}{3} = 2$
 $\bar{x} = \frac{4+3+2+2+4+6+2+1+3}{9} = 3$
 $SST = (4-3)^2 + (3-3)^2 + (2-3)^2 + (2-3)^2 + (4-3)^2 + (6-3)^2 + (2-3)^2 + (1-3)^2 + (3-3)^2$
 $= 1 + 0 + 1 + 1 + 1 + 9 + 1 + 4 + 0 = 18$
 $SSB = 3(3-3)^2 + 3(4-3)^2 + 3(2-3)^2$
 $= 0 + 3 + 3 = 6$
 $SSE = (4-3)^2 + (3-3)^2 + (2-3)^2 + (2-4)^2 + (4-4)^2 + (6-4)^2 + (2-2)^2 + (1-2)^2 + (3-2)^2$
 $= 1 + 0 + 1 + 4 + 0 + 4 + 0 + 1 + 1 = 12$

$H_0: \mu_1 = \mu_2 = \mu_3$
 $H_1: \mu_1 \neq \mu_2 \neq \mu_3$
 $SST = SST_{\text{treatment}} + S_{\text{error}}$
 $SST = SSB + SSE$

$S_{\text{error}}^2 = \frac{SSE}{n-1}$
 $n=9$
 $9-1=8$
 $2+2=4$
 $4+2=6$



15

The null hypothesis is I am going to assume that null hypothesis is $H_0: \mu_1 = \mu_2 = \mu_3$, alternative hypothesis is $\mu_1 \neq \mu_2 \neq \mu_3$. If you see even this technique name is called Anova analysis of variance. But I am comparing mean. What are you going to do here with the help of the concept called variance and I am going to compare the mean of three populations.

Nothing to do with, I am not going to, compare the variance and comparing the mean of three populations. So, far the group one have taken the sample mean that is 3 for group 2 I have taken sample means it is 4, group 3 it is 2. Then I find overall the sample mean $4 + 3 + 2 + 2 + 4 + 6 + 2 + 1 + 3$ with 9 elements, equal to 3. What I am going to do is I am going to find out the overall variance. That I am going to call it as SST, total sum of square.

Here, why I am saying it as variance, see the variance formula, we know that the variance what is the formula? Variance is equal to $(\Sigma X - \bar{X})^2$ Whole square divided by $n - 1$. This numerator that is $(\Sigma X - \bar{X})^2$ whole square that I am going to say sum of square. I am going to find out the overall variance that overall variance SST and I am going to group into two categories.

This variance is due to SS treatment plus variance, variance due to error. That error minus that their sum of square. So what I am writing SST is equal to total sum of square equal to some time there might be between columns SSB + SSE. So, obviously in the variance, due to treatment, that

is SSB is dominating we can see that the teaching methodology is influencing variable. First I am going to find out the overall variance.

For that variance I am going to find out only the numerator so that I am going to call it as SST. What is SST? How each element is away from overall mean. Overall mean is 3, so in the first column, the first element is 4. So, $4 - 3$ whole square the value in the second column is 3 . $3 - 3$ the whole square + $2 - 3$ whole square upto 18. So this 18 is nothing but total sum of square. This 18 I am going to see how much variance is due to this teaching methodology.

So I am going to call it is SSB. Some books call it SS treatment. That is treatment sum of square. What is treatment sum of square? In the first column, there are three element is there. 3 minus the first column mean is 3. This 3, so the overall mean is 3. So $3 (3 - 3)^2$ + the second column so this 3 represents the number of samples in each column, this 3 represents in the second column there are 3 elements.

So, this 4 represents mean of the second column is 4 minus 3 is overall mean 3 whole square + for third column also there are three elements. The mean of the third column is 2. 2 minus 3 this 3 is overall mean whole square. So, this becomes $0, 4 - 1, 1$ square into $3 = 3, 2 - 1 = 1$, is going to be 6. So this 6 is numerator of the variance that is SSB is 6. Then I will find out SSE. That is the variance error sum of square, the inherent variance.

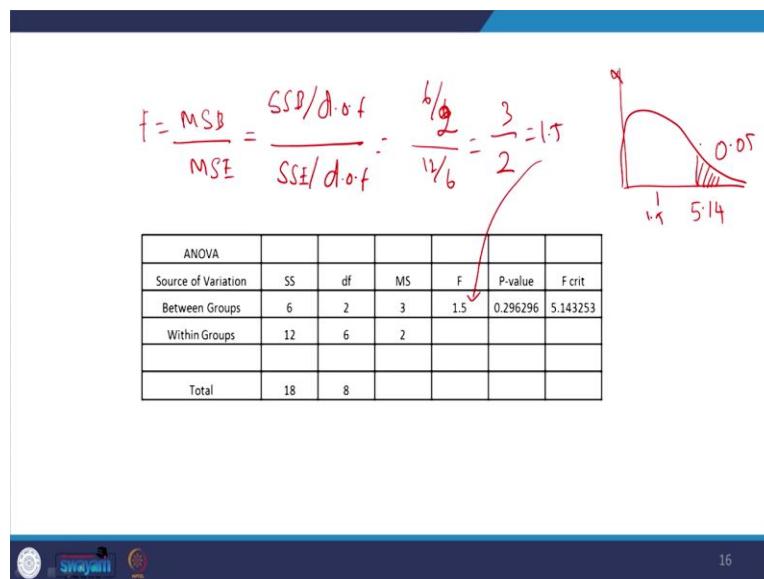
For inherent variance, when you look at the column 1 in the first element 4 the mean of the first column is 3. So, $4 - 3$ whole square + the second element is $3 - 3$ minus the mean of the first column is this 3 this this is a mean of first column. So, $4 - 3$ whole square + $(3 - 3)$ whole square $+(2 - 3)$ whole square. So this 4, this 4 represents mean of the second column at this represents this 2 represents the first element in the second column.

So, $2 - 4$ whole square + $4 - 4$ whole square + $6 - 4$ whole square. The third one is this 2 represents the mean of the third column. This 2 represent the first element in the third column $2 - 2$ whole square $+(1 - 2)$ whole square $+(3 - 2)$ the whole square. When you simplify that it is 12. So what happened the SST is divided into two categories one is variance due to treatment this

is Error variance or Individual variance. The second one is where to find out the degrees of freedom for SST.

What are the degrees of freedom? There are 9 elements. $9 - 1$ that is your degrees of freedom is 8. For SSB there 3 columns is there. So, 3 treatment. So, $3 - 1$ that is $3 - 1$ that is the degrees of freedom. For SSC, in the first column, there are three elements so the first column will have two degrees of freedom the second column the three elements. So, $3 - 1$, 2 degrees of freedom the third column also there are 3 elements. $3 - 1$, 2 degrees of freedom. So totally six degrees of freedom.

(Refer Slide Time: 19:32)



After that we have to find out f value. F value is nothing but MSB by MSE. That means what is MSB is mean square between columns. This is means error square. So what is this mean means, if you divide this SSB divided by the corresponding degrees of freedom we will get MSB. When you divide SSE divided by corresponding degrees of freedom we will get MSE. If we look at the previous one SSB, this is a two, degrees of freedom is 2. 6 divided by 2, SSE is how much chances?

That is 12. Degrees of freedom is 6, so 3 by 2 is 1.54 that 1.5 is nothing but this 1.5. Ok, Next what you have to do is, this is we can say your calculated value. You can refer your F table. In f table assume that alfa is equal to 5 %. You have to look at the, what is this value for example, if

you look at the table, it will give you 5.14 but your calculated value 1.5 is lying on the acceptance side table. This is F table so you have to accept null hypothesis.

This is an simple intuition for how this concept of Anova is working. In the next class we will do more theory behind this Anova, we will continue. Ok students in this class, we have seen how to find out the sample size for hypothesis testing. Then we have started the concept of Anova. For Anova I have taken one example then I have explained what is the SST, that is the total sum of square, treatment sum of square, error sum of square.

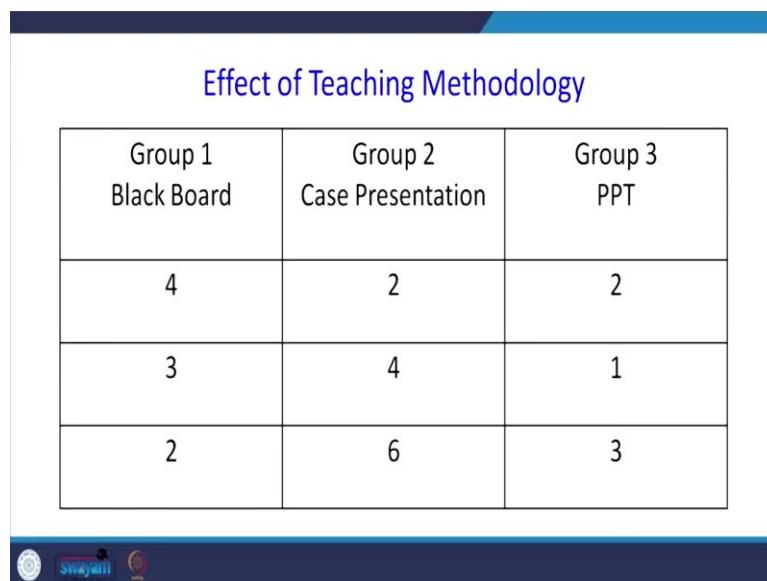
Then I have explained when should we go for Anova then I have solved one problem then, I will complete calculated F value with table F value then I have concluded the conclusion also. The next class, the same problem we will solve the help of python. Then, we will come then we'll go for post hoc analysis. Post hoc analysis is whenever you reject null hypothesis, we have to say which two pairs are different. So, that we will see in the next class, thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 24
ANOVA- II

Dear students in the previous class I have explained the concept behind analysis of variance. In this class with the help of Python will solve that problem because previously in the previous class we have solved it manually. Now we will use the help of Python will solve that problem this was the problem which was given.

(Refer Slide Time: 00:47)



Group 1 Black Board	Group 2 Case Presentation	Group 3 PPT
4	2	2
3	4	1
2	6	3

What is the problem is there are 3 different teaching methodology, we have to say which teaching methodology is more influencing on the student performance.

(Refer Slide Time: 01:01)

ANOVA with Python

```
In [15]: a=[4,3,2]
In [16]: b=[2,4,6]
In [17]: c=[2,1,3]
In [18]: stats.f_oneway(a,b,c)
Out[18]: F_onewayResult(statistic=1.5, pvalue=0.2962962962962962)
```

3

In Python so I have taken a is an array $a = [4, 3, 2]$ $b = [2, 4, 6]$ $c = [2, 1, 3]$ if you type `stats.f_oneway` just you call abc you run that one you will get this was were valuing this was for F that is a calculated F value, this is a p value. Suppose if the Alpha equal to 5 %, it is more than 5 % so we have to accept a null hypothesis that is what our previous result also.

(Refer Slide Time: 01:32)

Pandas.melt command

- `Pd.melt` allows you to ‘unpivot’ data from a ‘wide format’ into a ‘long format’, data with each row representing a data point.

4

Now we will use that another command that that is `pandas.melt` command we will see the purpose of this command for doing ANOVA `pd.melt` allows you to unpivot data from a wide format into long format that is data with each row representing a data point.

(Refer Slide Time: 01:51)

Jupyter code

```
In [22]: import pandas as pd  
import numpy as np  
import math  
from scipy import stats  
import scipy  
import statsmodels.api as sm  
from statsmodels.formula.api import ols  
from matplotlib import pyplot as plt
```

```
In [23]: data=pd.read_excel('oneway.xlsx')
```

```
In [24]: data
```

```
Out[24]:
```

	Teachin Method1	Teachin Method2	Teachin Method3
0	4	2	2
1	3	4	1
2	2	6	3

5

So, for that purpose input pandas dot pd import numpy is np, import math, from scipy import stats, import scipy, import statsmodel.api as sm, from statsmodels.formula.api import ols, from matplotlib import pyplot as plt, so first we will load the data the data I am going to save the data the given time in the excel file are going to save in the object called data. So, data = pd.read_excel ('oneway.xlsx').

So, I have loaded when I run this data now the data is appearing column 1 column 2 column 3 so 0 1 2 that is your index.

(Refer Slide Time: 02:42)

Transforming table

```
4]: data
```

```
4]:
```

	Teachin Method1	Teachin Method2	Teachin Method3
0	4	2	2
1	3	4	1
2	2	6	3



```
In [27]: data_new
```

```
Out[27]:
```

index	Treatments	value
0	0 Teachin Method1	4
1	1 Teachin Method1	3
2	2 Teachin Method1	2
3	0 Teachin Method2	2
4	1 Teachin Method2	4
5	2 Teachin Method2	6
6	0 Teachin Method3	2
7	1 Teachin Method3	1
8	2 Teachin Method3	3

7

Next what you have to do for running the data I need to have the data in this format what is that format is so T.M1 teaching methodology one teaching methodology if your teaching methodology one there may be some numbers. In teaching methodology 2 there may be some numbers and teaching methodology 3 there may be some numbers. So, this says your treatment so the next one says the value suppose I want to have the data in this format.

For that you have to use the following command that is data new I am going to call it is that way pd.melt(data.reset_index(), id_vars =['index'], value_vars = ['Teachin Method1', 'Teachin Method2', 'Teachin Method3']), data_new.Columns = [' index', 'treatment', 'value']. So, this is the this is the syntax for using the melt function.

So, if you run this data underscore new will get this kind of odd for do you see previously the data was 0 1 2 format now what we are saying all teaching methodology one it is grouped in this way this is group 1 this is your group 2 this is group 3. Now only one there was only 2 column one is a one is treatment another one is value. now after getting this data into this format for converting this purpose the melt command is used.

(Refer Slide Time: 04:22)

The screenshot shows a Jupyter Notebook interface with three code cells and their corresponding outputs:

- In [31]:
model=ols('value ~ C(Treatments)',data=data_new).fit()
- In [32]:
anova_table=sm.stats.anova_lm(model, typ=1)
- In [33]:
anova_table

Out[33]:

	df	sum_sq	mean_sq	F	PR(>F)
C(Treatments)	2.0	6.0	3.0	1.5	0.296296
Residual	6.0	12.0	2.0	NaN	NaN

So model equal to ols, ols is ordinary least square method in in quote value tilde C treatment, data equal to delta underscore new fit. Then if you write anova_table = sm.stats.anova_lm(model, typ = 1), when you run this will get the anova_table this represents

degrees of freedom for treatment because there was a 3 column treatment is 2 whereas dual is 6 because for a column 1 there are 3 elements so $3 - 1$, 2 degrees of freedom.

Similarly for column 2 also another 2 degrees of freedom for column 3 also another 2 degrees of freedom totally 6 degrees of freedom, so some squared is 6 here sum of square is 12. So, mean sum of square is 6 divided by 2 that is 6 this is 12 divided by 6 that is 2, so F value is 3 divided by 2 this was the p value this also we got it previously to help of; when you do it manually we compared this 1.5 and we got the same value this is with the help of Python we are also getting the same result.

(Refer Slide Time: 05:31)

Analysis of Variance: A Conceptual Overview

- Analysis of Variance (ANOVA) can be used to test for the equality of three or more population means
- Data obtained from observational or experimental studies can be used for the analysis
- We want to use the sample results to test the following hypotheses:
 $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$
 $H_a: \text{Not all population means are equal}$

9

Now we will go to the formal definition of ANOVA venire conceptual overview analysis of variance can be used to test the Equality of 3 or more population means. Data obtained from observational or experimental studies can be used for this analysis. We want to use the sampled result to test the following hypothesis in ANOVA what does the hypothesis $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$ it may be n number of columns.

(Refer Slide Time: 06:17)

Analysis of Variance: A Conceptual Overview

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

H_a : Not all population means are equal

- If H_0 is rejected, we cannot conclude that all population means are equal
- Rejecting H_0 means that at least two population means have different values

10

Alternative hypothesis not all population means are equal, H_0 equal to mu 1 equal to mu 2 equal to mu 3, H_a is not all population means are equal if it is not rejected we cannot conclude that all population means are equal so when you reject it that means there are some unusual means. So, rejecting H_0 means that at least two population means have different values.

(Refer Slide Time: 06:36)

Analysis of Variance: A Conceptual Overview

Assumptions for Analysis of Variance

- For each population, the response (dependent) variable is normally distributed
- The variance of the response variable, denoted σ^2 , is the same for all of the populations
- The observations must be independent

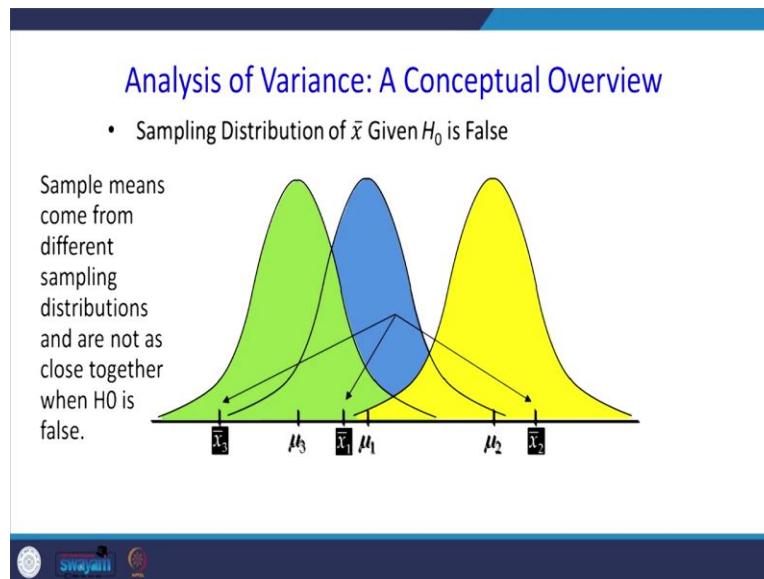
11

What are the Assumption for analysis of variance. For each population the response dependent variable is normally distributed. In our previous example the performance of the student is the dependent variable the independent variable is teaching methodology. The variance of the response variable denoted by Sigma square is the same for all populations. Why this assumption

is required? when you are comparing more than 2 groups the basic assumption is that the variance of that group should be same.

This concept we have explained when we are conducting 2 sample tests and the observation must be independent.

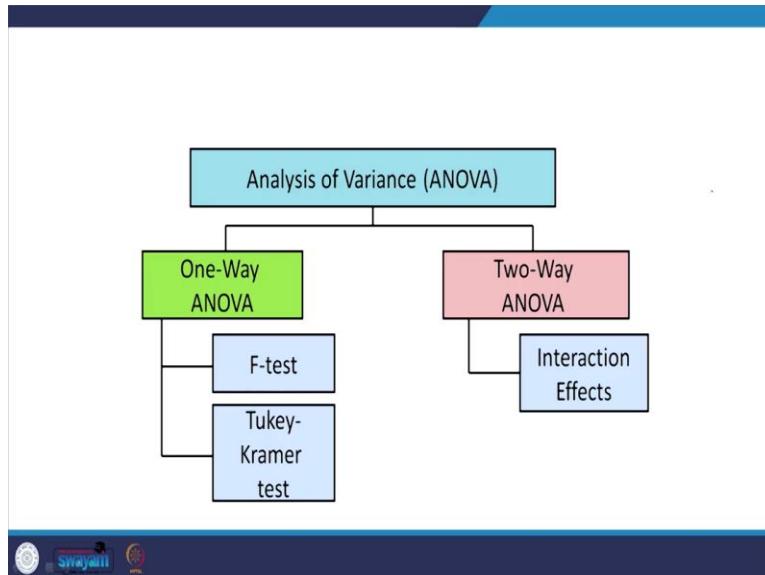
(Refer Slide Time: 07:16)



Look at this normal distribution this is the sampling distribution of \bar{x} given null hypothesis is a true sample means are close together because there is only one sampling distribution when H_0 is true. Look at this normal distribution there are 3 normal distribution here the sampling distribution of \bar{x} given H_0 is false what is H_0 is false what is H_0 here H_0 equal to μ_1 equal to μ_2 equal to μ_3 , if it is a false what will happen it would not be from same population it will be from different population.

So, the sample means come from different sampling distribution and are not as close together when H_0 is false.

(Refer Slide Time: 08:05)



In analysis of variance we can classify one is 1 way ANOVA another one is 2 way ANOVA there is one more thing in between that is called R B D randomized block design will see when we will go for randomized block design. In one-way ANOVA we are going to do the F test the F test will help you to decide whether the null hypothesis accepted or rejected when you reject the null hypothesis then Tukey Kramer test will help you which 2 pairs are equal or which 2 pairs are not equal.

Then this side is a 2-way ANOVA then we will go for interaction effects that we will see in coming classes. In this class we will see how to do the F test how will you do the 2 Tukey Kramer test.

(Refer Slide Time: 08:49)

General ANOVA Setting

- Investigator controls one or more factors of interest
 - Each factor contains two or more levels
 - Levels can be numerical or categorical
 - Different levels produce different groups
 - Think of the groups as populations
- Observe effects on the dependent variable
 - Are the groups the same?
- Experimental design: the plan used to collect the data



What is the general ANOVA setting investigator controls one or more factors of interest in our previous example the teaching methodology is the factor. So, each factor contains 2 or more level in our case you see suppose of the pressure is the one parameter may be high or low that is level. So, high is one level low is another level, level can be numerical or categorical. Here the example is it is a categorical he need not be categorical it may be a continuous variable also.

So, different levels produce different groups think of the groups as population we can say each groups can be we can consider as the population. Observe effect on the dependent variable so what we are going to doing otherwise what is the effect of this treatment on the dependent variable. Next we will see experimental design the plan used to collect the data only the external design will have a plan to collect the data. And will see the effect of this data on the how this treatment is influencing the data.

(Refer Slide Time: 09:59)

Completely Randomized Design

- Experimental units (subjects) are assigned randomly to the different levels (groups)
 - Subjects are assumed homogeneous
- Only one factor or independent variable
 - With two or more levels (groups)
- Analyzed by one-factor analysis of variance (one-way ANOVA)



The first one method called completely randomized design in our previous example the students are allocated to 3 groups randomly that is an example of you were completely randomized design. There is no bias because what will happen there suppose if you consider the student IQ level then you are allocating that is student to different category of classes then that is not called biased method. So, what is happening here the experimental units are assigned randomly to the different levels so subjects are assumed homogeneous.

Only one factor or one independent variable that is called one way ANOVA because here teaching methodology the independent variable the student performance that is the marks is the dependent variable. There also we can have 2 or more levels if you are analyzing one factor analysis of variance it is called one way one way ANOVA. If there are 2 independent variable that is a 2 way ANOVA.

(Refer Slide Time: 10:57)

Analysis of Variance and the Completely Randomized Design

- Between-Treatments Estimate of Population Variance
- Within-Treatments Estimate of Population Variance
- Comparing the Variance Estimates: The *F* Test
- ANOVA Table

17

So, what we are doing the basic concept behind is we are finding the variance due to between the treatment and variance between the treatments. So, that will go to your numerator that is nothing but every SS treatment that is why I wrote SSB. When you divide by degrees of freedom this is MSB, divided by the SSE divided by degrees of freedom that is variance within the treatment this is nothing but you are MSE. So, we will find variance between the treatment then variance within the treatment then we will go for F test then I will explain what is this ANOVA table.

(Refer Slide Time: 11:44)

Analysis of Variance and the Completely Randomized Design

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

$$H_a: \text{Not all population means are equal}$$

- Assume that a simple random sample of size n_j has been selected from each of the k populations or treatments. For the resulting sample data, let

x_{ij} = value of observation i for treatment j

n_j = number of observations for treatment j

\bar{x}_j = sample mean for treatment j

s_j^2 = sample variance for treatment j

s_j = sample standard deviation for treatment j

19

So, what is a null hypothesis for the CRD, completely randomized design μ_1 equal to μ_2 = μ_3 equal to μ_k not all population means are equal. Here assume that a simple random sample of n_j has been selected from each of the k populations or treatment there are k treatment in our

previous example there was a 3 treatment 3 factors 3 factors means 3 levels for the resulting sample data let X_{ij} value of observation i for treatment j , n_j is number of observation for treatment j , \bar{x}_j is sample mean for treatment say, s_j^2 square sample variance for treatment j , s_j is the sample standard deviation for treatment j

(Refer Slide Time: 12:32)

Between-Treatments Estimate of Population Variance σ^2

- The estimate of σ^2 based on the variation of the sample means is called the mean square due to treatments and is denoted by MSTR

$$\text{MSTR} = \frac{\sum_{j=1}^k n_j (\bar{x}_j - \bar{\bar{x}})^2}{k-1}$$

Denominator is the degrees of freedom associated with SSTR

Numerator is called the sum of squares due to treatments (SSTR)


Swayam

GATE
20

First we will find out between treatment estimation of population variance Sigma square. The estimate of Sigma square based on the variation of the sample mean is called mean square due to treatment that is denoted by MSTR in our example previously we have used we can have MSB that is mean square due to between columns. So, how we are finding is this MSTR is nothing but n_j number of elements in column j \bar{x}_j that column j mean minus the overall mean whole square divided by $k - 1$, k is the number of columns here $3 - 1$.

So the denominator is the degrees of freedom associated with sum of square treatment the numerator is called the sum of square due to treatment SSTR divided by degrees of freedom.

(Refer Slide Time: 13:34)

Between-Treatments Estimate of Population Variance σ^2

- Mean Square due to Treatments (MSTR)

$$MSTR = \frac{\sum_{j=1}^k n_j (\bar{x}_j - \bar{\bar{x}})^2}{k-1}$$

Where:

k = number of groups

n_j = sample size from group j

\bar{x}_j = sample mean from group j

$\bar{\bar{x}}$ = grand mean (mean of all data values)



21

Between treatment estimation of population variance Sigma square the mean square due to treatment that formula which you have seen previous slides so what is the meaning of this k , k is the number of groups, k number of columns n_j is the sample size from Group j x_j bar is the sample mean from Group j x double bar is the grand mean, mean of all data values over all mean.

(Refer Slide Time: 13:58)

Within-Treatments Estimate of Population Variance σ^2

- The estimate of σ^2 based on the variation of the sample observations within each sample is called the mean square error and is denoted by MSE

$$MSE = \frac{\sum_{j=1}^k (n_j - 1)s_j^2}{n_T - k}$$

Denominator is the
degrees of freedom
associated with SSE

Numerator is called
the sum of squares
due to error (SSE)



22

Next we will see within treatment estimate our population variance. The estimator of Sigma square based on the variation of the sample observations within each sample this is more important term within each sample is called mean squared error is denoted by MSE. So, mean

square error how we are doing that one n_j in column j how many element is there minus one, that is our degrees of freedom s_j square.

Actually how it has come you see if we want to know s_j square what is a formula $\sum(X - \bar{X})^2$ divided by $n - 1$, so this instead of writing numerator that can be written as $(s_j)^2 \cdot (n - 1)$ that is why it is written $n_j - 1$ $(s_j)^2$ or n_T is denominator is the degrees of freedom associated with error sum of square I will tell you that.

What is the n_T in the next slide K is the number of groups n_T is the number of treatment here this n_T is nothing but the overall degrees of freedom. From the overall degrees of freedom if you subtract the degrees of freedom for between the columns then you will get either degrees of freedom for your SSE that is error sum of square. In our previous example we might have seen the n_T is $9 - 1$, 8 and the K is there was 3 columns, so it is 2 it was 6 degrees of freedom in our previous problem for MSE.

(Refer Slide Time: 15:43)

Within-Treatments Estimate of Population Variance σ^2

- Mean Square Error (MSE)

$$MSE = \frac{\sum_{j=1}^k (n_j - 1)s_j^2}{n_T - k}$$

Where:

k = number of groups
 n_j = number of observations for treatment j
 s_j^2 = sample variance for treatment j

23

The formula for mean squared error is \sum of j equal to 1 to k , $(n_j - 1) (s_j)^2$ by $(n_T - k)$ where n_T is total number of observations where k is number of groups.

(Refer Slide Time: 16:00)

Comparing the Variance Estimates: The *F* Test

- If the null hypothesis is true and the ANOVA assumptions are valid, the sampling distribution of MSTR/MSE is an *F* distribution with MSTR d.f. equal to $k - 1$ and MSE d.f. equal to $n_T - k$.
- If the means of the k populations are not equal, the value of MSTR/MSE will be inflated because MSTR overestimates σ^2
- Hence, we will reject H_0 if the resulting value of MSTR/MSE appears to be too large to have been selected at random from the appropriate *F* distribution



24

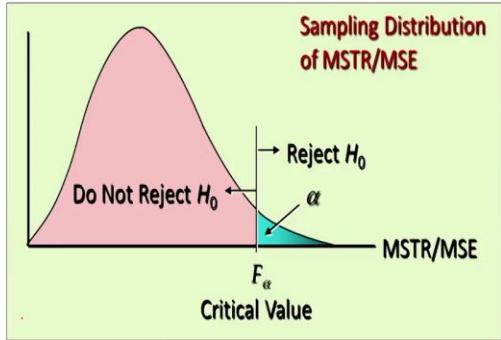
Comparing the variance estimates that is F test if the null hypotheses are true and ANOVA assumptions are valid the sampling distribution of MST are divided by MSE is an *F* distribution with MSTR degrees of freedom is equal to $k - 1$ that is number of column minus 1 and MSE a degrees of freedom is $n_T - k$, n_T is total number of sample minus k is number of groups. If the means of the K populations are not equal the value of MSTR divided by MSE will be inflated because MSTR overestimate Sigma square.

So, what is the meaning of this one is we are finding you have the value of *F* is MSTR divided by MSE there are 2 possibility it may be equal 1 or less than 1 or greater than 1. If it is equal to 1 what is the meaning is variance due to treatment is equal to variance due to individual error. If it is greater than 1 the variance due to treatment is more when compared to within the error. When it become less than 1 if the MSE that is error due to individual difference is more when compared to treatment then it will become less than 1.

So, you see this one if the mean of the k populations are not equal the value of MSTR by MSE will be inflated because the MSTR overestimate Sigma square. Hence we will reject because *F* value become very big when *F* value is very big we will reject H_0 if the resulting value of MSTR by MSE appears to be too large to have been selected at random from appropriate *F* distribution.

(Refer Slide Time: 17:52)

Comparing the Variance Estimates: The F Test



This is situation so what will happen when F is bigger number or obviously will be landing on the rejection site will reject null hypothesis. When you reject a null hypothesis we will say $\mu_1 = \mu_2 = \mu_3$ this was your null hypothesis, alternative hypothesis is μ_1 not equal to μ_2 not equal to μ_3 . So, when you reject null hypothesis we can conclude that these means are not equal and one more thing this is the F distribution it is not normal distribution it is a right skewed distribution.

(Refer Slide Time: 18:35)

ANOVA Table for a Completely Randomized Design

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-Value
Treatments	SSTR	$k - 1$	$MSTR = \frac{SSTR}{k-1}$	$\frac{MSTR}{MSE}$	
Error	SSE	$n_T - k$	$MSE = \frac{SSE}{n_T - k}$		
Total	SST	$n_T - 1$			

SST is partitioned into SSTR and SSE.

SST's degrees of freedom (d.f.) are partitioned into SSTR's d.f. and SSE's d.f.

This is the ANOVA table setup so what will be written sources of variation. So, there may be variation may be due to treatment variance due to error, so sum of square is sum of square treatment error sum of square. Here the deal is of freedom is $K - 1$ here $n_T - k$ generally if you

SST degrees of freedom is $n_T - 1$ and n_T is total number of elements - 1 when you subtract this $n^2 - 1 - k - 1$ that will give you $n_T - K$ so MSTR is nothing but we have to divide this SSTR a little bit corresponding degrees of freedom. so, it will become mean treatment sum of square.

When you divide by SSE you total by corresponding degrees of freedom mean error sum square. So, the ratio of you see a MSTR divided by MSE always in the denominator there should be error term because when you go for 2-way ANOVA be able to remember that the denominator always there will be error term then we can find out corresponding p value this is what we have done previously when we are explaining my first example.

(Refer Slide Time: 19:41)

ANOVA Table for a Completely Randomized Design

- SST divided by its degrees of freedom $n_T - 1$ is the overall sample variance that would be obtained if we treated the entire set of observations as one data set.
- With the entire data set as one sample, the formula for computing the total sum of squares, SST, is:

$$SST = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 = SSTR + SSE$$

27

Generally what is happening this SST divided by its degrees of freedom $nT - 1$ is the overall sample variance that would be obtained if you treated the entire setup observation as one data set right when you divide this SST by corresponding degrees of freedom that is overall variance. With entire data set as a one sample the formula for computing the total sum of square is SST is $\sum j$ equal to 1 to $k \sum i$ equal to 1 to n say $(X_{ij} - X_{\text{double bar}})^2$.

So, this total sum of square can be splitted into 2 part one is treatment sum of square and error sum of square. If this treatment sum of square is dominating even without going further test we can say that there is a influence of treatment on the response variable.

(Refer Slide Time: 20:35)

ANOVA Table for a Completely Randomized Design

- ANOVA can be viewed as the process of partitioning the total sum of squares and the degrees of freedom into their corresponding sources: treatments and error
- Dividing the sum of squares by the appropriate degrees of freedom provides the variance estimates and the F value used to test the hypothesis of equal population means.

28

ANOVA can be viewed as the process of partitioning the total sum of square and the degrees of freedom into their corresponding sources that is treatment and error. Dividing the sum of square by the appropriate degrees of freedom provides the variance estimates and the F value used to test the hypothesis of equal population means.

(Refer Slide Time: 21:02)

Test for the Equality of k Population Means

- Hypotheses

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

H_a : Not all population means are equal

- Test Statistic

$$F = \frac{MSTR}{MSE}$$

29

What is the hypothesis the null hypothesis as usual mu1 equal to mu2 equal to mu 3, alternative hypothesis: not all population means are equal the test statistic is the ratio of mean treatment sum of square divided by mean error sum of square.

(Refer Slide Time: 21:19)

Test for the Equality of k Population Means

p- Value Approach

Critical Value Approach

Reject H_0 if $p\text{-value} \leq \alpha$

Reject H_0 if $F \geq F_\alpha$

Where the value of F_α is based on an F distribution with $k - 1$ numerator d.f. and $n_T - k$ denominator d.f.



30

The p-value approach as usual for hypothesis testing also if the p-value is less than or equal to alpha we have to reject a null hypothesis. If you are using critical value the F value is greater than your value which you got from the table that also we have to reject our null hypothesis. In this class what we have seen we have taken one problem that is problem we have solved with help of Python then I have explained the theoretical background behind this ANOVA.

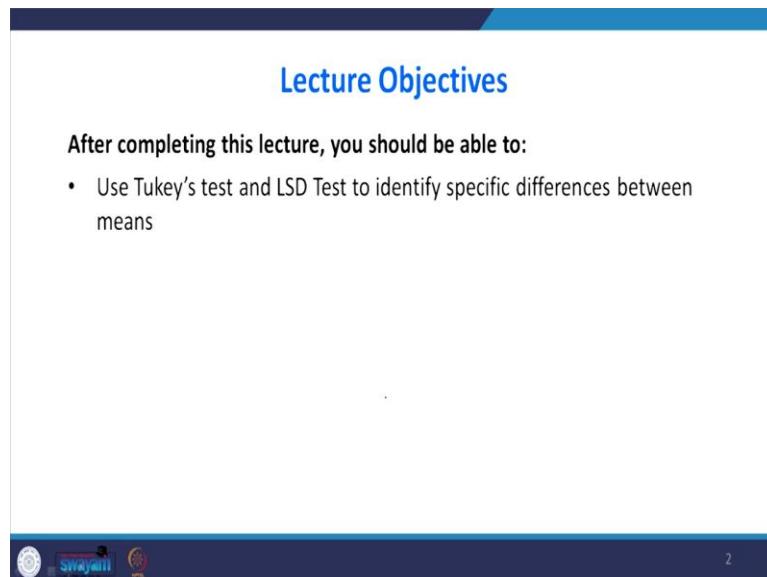
Then I have explained what is the total sum of square then what is the treatment sum of square than error sum of square. Then what is the degrees of corresponding degrees of freedom. In the next class will take extension of these classes once we reject a null hypothesis we have to say which 2 means are equal or not equal. So, that analysis is post hoc analysis we will continue the next lecture with the new topic of post hoc analysis in ANOVA, thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 25
Post Hoc Analysis(Tukey's test)

Dear students in the previous class we have seen the theoretical background behind analysis of variance and also we have solved a problem. So, what will happen once you reject a null hypothesis of an ANOVA problem that means you are accepting alternative hypothesis it is need not be that all means not equal some time any pair may be equal some other pair may be not equal. So, whenever you reject a null hypothesis then we have to say which two means are equal for that purpose there is one more statistical analysis that is a Post Hoc analysis useful one test is Tukey Kramer test another test is HSD test. We will see with that how to use this post hoc analysis in ANOVA in this lecture.

(Refer Slide Time: 01:22)



The slide has a blue header bar and a blue footer bar. The main content area is white. The title 'Lecture Objectives' is centered in blue text. Below it, a statement in black text says 'After completing this lecture, you should be able to:' followed by a bulleted list. The footer bar contains three small icons on the left and the number '2' on the right.

Lecture Objectives

After completing this lecture, you should be able to:

- Use Tukey's test and LSD Test to identify specific differences between means

So, the lecture objective is after completing the lecture you should be able to use Tukey test and least that is LSD test to identify specific differences between means.

(Refer Slide Time: 01:34)

Designing engineering experiments

- Experimental design methods are also useful in engineering design activities, where new products are developed and existing ones are improved
- By using designed experiments, engineers can determine which subset of the process variables has the greatest influence on process performance

3

We will take in this problem and engineering perspective experimental design methods also useful in engineering design activities where new products are developed and existing ones are improved. By you see design of experiments engineers can determine which subset of the process variables has the greatest influence on the process performance that is the main objective of the design of experiment is. What kind of variables that has the greatest influence on the performance of the product.

(Refer Slide Time: 02:09)

Designing engineering experiments

- The results of an experiment can lead to
 1. Improved process yield
 2. Reduced variability in the process and closer conformance to nominal or target requirements
 3. Reduced design and development time
 4. Reduced cost of operation

4

When you do the design of experiments what are the advantages benefits one is it improves the process yield it reduces the variability in the process that leads to less rejection and closer confirmation to the nominal or target requirements, so quality of the product is improved and

reduced design and development time because before making the product since we are doing the design experiments so the time spent on redesign is reduced at the same time reduced the cost of operations because the waste is minimized.

(Refer Slide Time: 02:48)

Designing engineering experiments

- Every experiment involves a sequence of activities:
 1. **Conjecture**—the original hypothesis that motivates the experiment
 2. **Experiment**—the test performed to investigate the conjecture
 3. **Analysis**—the statistical analysis of the data from the experiment
 4. **Conclusion**—what has been learned about the original conjecture from the experiment. Often the experiment will lead to a revised conjecture, and a new experiment, and so forth

Some of the term that we have to remember while going for this design of experiment is one does a conjecture, Conjecture is the original hypothesis that motivates the experiment, experiment the tests performed to investigate the conjecture. Analysis the statistical analysis of the data from the experiment, so, conclusion is what has been learned about the original conjecture from the experiment often the experiment will lead to a revised conjecture and new experiment and so forth.

(Refer Slide Time: 03:23)

The completely randomized single-factor experiment example

- A manufacturer of paper that is used for making grocery bags is interested in improving the tensile strength of the product
- Product engineer thinks that tensile strength is a function of the hardwood concentration in the pulp and that the range of hardwood concentrations of practical interest is between 5 and 20%.



Reference: Applied statistics and probability for engineers, Douglas C. Montgomery, George C. Runger, John Wiley & Sons, 2007



6

We will solve one, one way problem in this class then I will explain how to use post hoc analysis. The problem is like this a manufacturer of paper industry he is using the paper for making grocery bags and he want to improve the tensile strength of the product. The product engineer thinks that the tensile strength is a function of hardwood concentration in the pulp and that the range of hardwood concentration is concentration of practical interest is between 5 to 20% so what is the meaning is that when you increase the hardwood concentration so the tensile strength will increase.

A team of engineers responsible for the study decides to investigate for level of hardwood concentrations the concentration level which they consider our 5% , 10% 15% and 20% they decide to make up 6 test specimen at each concentration level using a pilot plant all 24 specimens are tested on a laboratory tensile tester in a random order the data from this experiment is shown in the table.

(Refer Slide Time: 04:40)

The completely randomized single-factor experiment example

- Tensile Strength of Paper (psi)

Hardwood Concentration (%)	Observations						Total	Avg
	1	2	3	4	5	6		
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96



8

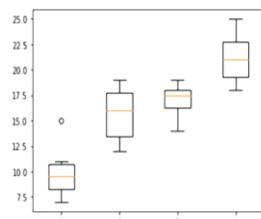
Now the row 5 represents hardwood concentration 5% 10% 15% and 20% 5% 10% these are the observations there are 6 times the experiment is repeated the average value is given so here the treatment is the percentage of hardwood concentration.

(Refer Slide Time: 05:02)

The completely randomized single-factor experiment example

```
In [3]: fivepercent=[7,8,15,11,9,10]
tenpercent=[12,17,13,18,19,15]
fifteenpercent=[14,18,19,17,16,18]
twentypercent=[19,25,22,23,18,20]

box_plot_data=[fivepercent,tenpercent,fifteenpercent,twentypercent]
plt.boxplot(box_plot_data)
plt.show()
```



9

First we will plot it with help of box and whisker plot so I am going to take the first data set as a 5% in array 5 8 15 11 9 10, next variable name is a 10 % I am ever taken the next array 12, 7, 13, 18, 19, 15 then 15% and I have taken all the 6 variables in 15% in 20% and so on. So, I go to draw the box plot, so box underscore plot underscore data so just to call that arrays for 5% 10% 15% and 20% then you write plt. Boxplot(box_plot_data), plt.show, so we are getting the box and whisker plot.

So what is happening here you see the means are not equal there is a lot of differences they there appears that whenever the percentage of hardwood is increasing so the tensile strength is increasing because there seems to be there is an increasing strength.

(Refer Slide Time: 06:13)

Treatment	Observations				Totals	Averages
1	y_{11}	y_{12}	...	y_{1n}	$y_{1..}$	$\bar{y}_{1..}$
2	y_{21}	y_{23}	...	y_{2n}	$y_{2..}$	$\bar{y}_{2..}$
.
.
.
a	y_{a1}	y_{a2}	...	y_{an}	$y_{a..}$	$\bar{y}_{a..}$
					$y_{...}$	$\bar{y}_{...}$

This is a typical data for single factor experiment generally see the treatment is taken as 1 2 3 4 observations are taken in row wise y_{11} this is the response variable of first a treatment and first a sample first treatment second sample first treatments and y_{1n} so if I write $y_{1..}$ that is a total the first row total $Y_{1..}$ second row total. If I write a is there is 'a' treatment 'a' levels is it so if I write $Y_{a..}$ represents the sum so $Y_{a..}$ means the totals.

If I write $\bar{Y}_{1..}$ that is the row 1 mean $\bar{Y}_{2..}$ that is the second row mean by a dot at level averages. If you write $\bar{Y}_{..}$ it is the sum of all total if I write $\bar{y}_{...}$ that is the average.

(Refer Slide Time: 07:17)

Sum of Squares

$$\text{Total sum of squares} = SS_T = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2$$

$$\text{Treatment sum of squares} = SS_{\text{Treatments}} = n \sum_{i=1}^a (\bar{y}_i - \bar{y}_{..})^2$$

$$\text{Error sum of Squares} = SSE = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_j)^2$$



So, SST treatment sum of square is the formula which you have seen previously because if you write in this way for especially for as a student this will it is easy way to solve the problem in the examination. So, this will save a lot of time Sigma I equal to 1 to a up to level j equal to 1 to n up to number of sample size y ij is the individual response - y dot dot but that is overall mean so that will give the total sum of square.

If you if you want to move the treatment sum of square SS treatment equal to n into $y_i - \bar{y}_{..}$ that is over all this is the row 1 mean this is overall mean whole square multiplied by n number of sample in the row 1 y. SSE is $y_{ij} - \bar{y}_j$ that corresponding mean whole square okay this is your error sum of square.

(Refer Slide Time: 08:25)

ANOVA with Equal Sample Sizes

$$SST = \sum_{i=1}^a \sum_{j=1}^n y_{ij}^2 - \frac{y_{..}^2}{N}$$

$$SSTreatments = \frac{1}{n} \sum_{i=1}^a y_i^2 - \frac{y_{..}^2}{N}$$

N = an = No. of Treatments x no. of sample size = Total no. of Sample Size



There is a shortcut formula for this the shortcut formula because this formula is very useful if you are using calculator. So, when you simplify the previous slide with what you are done so SST is nothing but Sigma of I equal to 1 to a sigma j equal to 1 to n y_{ij} square - y dot dot squared order by n, n is the nothing but a into n, a is the number of level n is small n is number of observation at level a. So, trick this is total sum of square the treatment sum of square is 1 by n Sigma i equal to 1 to a y_i dot m square - y dot dot whole square.

When you see that here also y dot dot square by n here also y dot square n is same so if you calculate for 1 it can be used for both calculation. So, we know that SST equal to SSTreatment plus SSE so when you subtract it you can get SSE, so this will save a lot of time in the examination. The previously there is a equal sample size if there is unequal sample size SST is same y_{IJ} square - y dot square by n but the SS Treatment is y_i dot squared order by n i because this this new term will be introduced there - y double dot whole square divided by n.

(Refer Slide Time: 09:51)

Problem: Analysis of variance

- Consider the paper tensile strength experiment described.
- We can use the analysis of variance to test the hypothesis that different hardwood concentrations do not affect the mean tensile strength of the paper.
- The hypotheses are
- $H_0: \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$
- $H_1: \tau_i \neq 0$ for at least one i

14

Consider the paper tensile strength experiment which is described previously. We can use the analysis of variance to test hypothesis that the different hardwood concentration do not affect the mean tensile strength of the paper. So, what is the null hypothesis is that that different hardwood concentration does not affect the mean tensile strength that means the hardwood concentration tensile strength are independent nothing to do with that one.

So, here the hypothesis is different treatment effect is 0, the alternative hypothesis of this is there is an effect of treatment.

(Refer Slide Time: 10:30)

Problem: Analysis of variance

- We will use $\alpha = 0.01$.
- The sums of squares for the analysis of variance are computed are as follows:

$$\begin{aligned} SS_T &= \sum_{i=1}^4 \sum_{j=1}^6 y_{ij}^2 - \frac{\bar{y}_{..}^2}{N} \\ &= (7)^2 + (8)^2 + \dots + (20)^2 - \frac{(383)^2}{24} = 512.96 \\ SS_{\text{Treatments}} &= \sum_{i=1}^4 \frac{\bar{y}_{i.}^2}{n} - \frac{\bar{y}_{..}^2}{N} \\ &= \frac{(60)^2 + (94)^2 + (102)^2 + (127)^2}{6} - \frac{(383)^2}{24} = 382.79 \\ SS_E &= SS_T - SS_{\text{Treatments}} \\ &= 512.96 - 382.79 = 130.17 \end{aligned}$$

15

When we take alpha equal to 1% the sum of square of analysis of variance are computed as follows we can say $y_{ij}^2 - (y_{..}^2 / N)$, we can simplify we just you can substitute this formula so SS treatment is 512.96, sorry total sum of square SS treatment is 382.79 when you subtract it will get SSE so SSE is 512.96 – 382.79 it is 130.17.

(Refer Slide Time: 11:01)

Problem: Analysis of variance

- The ANOVA is summarized as follow

Source of Variation	Sum of Squares	Degrees of freedom	Mean Square	F ₀	P-value
Hardwood concentration	382.79	3	127.6	19.6	3.59 E-6
Error	130.17	20	6.51		
Total	512.96	23			

In [9]: from scipy import stats
1-scipy.stats.f.cdf(19.6, 3, 20)
Out[9]: 3.599599239012541e-06

17

So, this is an ANOVA setup when you supply this value here so you can get what you are done we got SST we got SS treatment when you subtract it will get SSE error so degrees of freedom is there is a totally 24 element so $24 - 1$, 23. So, there was a 4 rows, $4 - 1$, 3 so $23 - 3$ is 20 then when you divide this 382.9 degrees of freedom we will get 127.6 when you divide 172 divided by 20 will get 6.5. so, 127.62 by 6.5 you will be 19.6 so it look like 19.6 big, so what we can do so it said it will this is a calculated value.

So what is a table value so when F is 19.6 numerator degrees of freedom is 3 denominator degrees of freedom is 20, so when you subtract it so 1 minus so that will give you the p value p value is 3.59 into 10 to the power - 6 you see this values is very, very, very low. So, we are two because this p-value is less than alpha if you say alpha equal to 5% we have to reject null hypothesis when you reject a null hypothesis that there is the influence of this hardwood on the tensile strength.

(Refer Slide Time: 12:37)

Problem: Analysis of variance

- Since $F_{0.01,3,20} = 4.94$, we reject H_0 and conclude that hardwood concentration in the pulp significantly affects the mean strength of the paper

```
In [32]: scipy.stats.f.ppf(1-0.01, dfn=3, dfd=20)
```

```
Out[32]: 4.938193382310539
```

18

Since $F_{1\% | 3, 20}$ is 4.94 so when you $| 3, 20$ you see that when you suppose if you are comparing the critical value so this value is 4.9 which you got from this one you can see that `scipy.stats.f.ppf`, if it is a 1% so we want to know probability of when the right side area is 0.01 so for that `scipy.stats.f.ppf(1 - 0.01)` will give you the because we want to know the right side area but the Python gives only the left side area so $1 - 0.01$ that probability when degrees of freedom is 3 when the denominator is no 20 the corresponding F value is 4.93. Our calculated F value is 19.6, so 19.6 far away from here so we got to reject a null hypothesis.

(Refer Slide Time: 13:43)

Problem: Analysis of variance

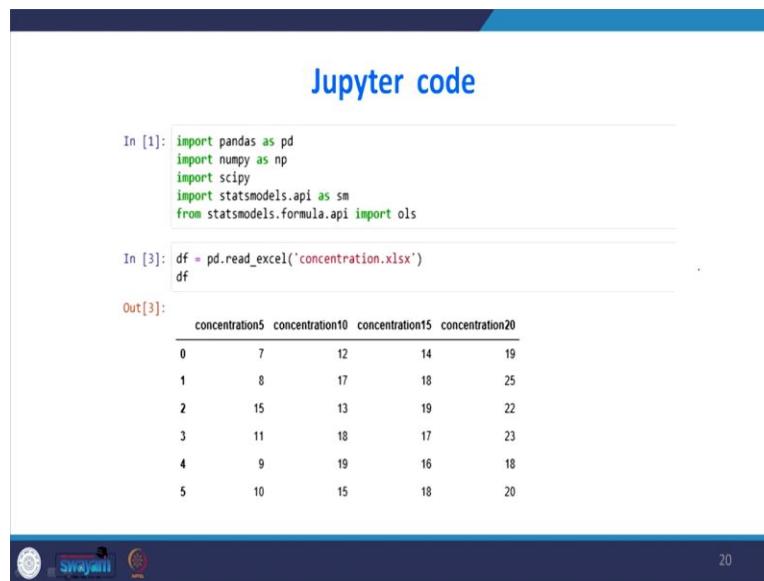
```
In [23]: scipy.stats.f_oneway(fivepercent,tenpercent,fifteenpercent,twentypercent)
```

```
Out[23]: F_onewayResult(statistic=19.605206999573184, pvalue=3.5925782584743027e-06)
```

19

So, what we can do we can directly we can run ANOVA so `scipy.stats.f_oneway`, call 5% 10% 15% 20% so you are getting F statistics 19.68 CP value is 3.59 into 10 to power – 6, so both the way you are getting the same result.

(Refer Slide Time: 14:06)



Jupyter code

```
In [1]: import pandas as pd
import numpy as np
import scipy
import statsmodels.api as sm
from statsmodels.formula.api import ols

In [3]: df = pd.read_excel('concentration.xlsx')
df
```

Out[3]:

	concentration5	concentration10	concentration15	concentration20
0	7	12	14	19
1	8	17	18	25
2	15	13	19	22
3	11	18	17	23
4	9	19	16	18
5	10	15	18	20

Now we will solve this problem suppose assume that this dataset which I have already entered into the excel file. So, import pandas as pd, import numpy as np, import scipy import statsmodels.api as sm, from statsmodels.formula .api import ols, so `df = pd.read_excel` I saved that file name is `concentration.xlsx`. So, when you write this `df` you are getting so concentration 1 concentration 2 concentration 3 consultation for that is a different % level.

Now we have to convert this one into only in two column in one column I need to have concentration level in our another column I am going to have only the values response variables.

(Refer Slide Time: 14:56)

Jupyter code

```
In [5]: data_r1 = pd.melt(d.reset_index(), id_vars=['index'], value_vars=['concentration5','concentration10','concentration15','concentration20'])
data_r1.columns = ['index', 'treatments', 'value']

In [6]: model = ols('value ~ C(treatments)', data=data_r1).fit()

In [7]: model.summary()
```

21

For that I have to use melt function so I am going to save that file in the object called data underscore r 1, pd . melt d f dot reset underscore index, id underscore vars equal to index, value underscore vrs equal to concentrations file that is the value of the variables concentration of 5 concentration 10 concentration 15 concentration 20 so data underscore r 1 columns going to be index treatment and values. So, model ols values tilda c in bracket treatments data equal to data underscore r 1 dot fit. So when I write model dot summary I will be getting this result right.

(Refer Slide Time: 15:44)

Jupyter code

Out[7]: OLS Regression Results

Dep. Variable:	value	R-squared:	0.746		
Model:	OLS	Adj. R-squared:	0.708		
Method:	Least Squares	F-statistic:	19.61		
Date:	Tue, 27 Aug 2019	Prob (F-statistic):	3.59e-06		
Time:	15:03:38	Log-Likelihood:	-54.344		
No. Observations:	24	AIC:	116.7		
Df Residuals:	20	BIC:	121.4		
Df Model:	3				
Covariance Type:	nonrobust				
	coef	std err	t	P> t	[0.025 0.975]
Intercept	15.6687	1.041	15.042	0.000	13.494 17.839
C(treatments)[T.concentration15]	1.3333	1.473	0.905	0.376	-1.739 4.406
C(treatments)[T.concentration20]	5.5000	1.473	3.734	0.001	2.428 8.572
C(treatments)[T.concentration5]	-5.6667	1.473	-3.847	0.001	-8.739 -2.594
Omnibus:	0.929	Durbin-Watson:	2.181		
Prob(Omnibus):	0.628	Jarque-Bera (JB):	0.061		
Skew:	0.248	Prob(JB):	0.650		
Kurtosis:	2.215	Cond. No.:	4.79		

22

So what we are getting here this was the regression model now we are getting the ANOVA table so aov_table = sm.stats.anova_lm when you call that model, typ = 1, you type this table you will get treatment there are 4 row was there so 4 - 1 3 degrees of freedom for the error degrees of

freedom is 20 so this is treatment sum of square error sum of square when you divide by 3 you will get 127.59 when you divide 130 by 20, 6.59 so when you divide 127.59 into 509 this is your calculated values the p-value is very, very low. So, we can reject the null hypothesis.

(Refer Slide Time: 16:37)

Multiple Comparisons Following the ANOVA

- When the null hypothesis is rejected in the ANOVA, we know that some of the treatment or factor level means are different
- ANOVA doesn't identify which means are different
- Methods for investigating this issue are called multiple comparisons methods



24

When null hypothesis rejected in the ANOVA we know that sum of the treatment our factor level means are different. Anova does not identify which means are different methods for investigating this issue is called multiple comparison method are post hoc analysis that we will do here.

(Refer Slide Time: 16:57)

Fisher's least significant difference (LSD) method

- The Fisher LSD method compares all pairs of means with the null hypotheses $H_0: \mu_i = \mu_j$ (for all $i \neq j$) using the t-statistic

$$t_0 = \frac{\bar{y}_i^* - \bar{y}_j^*}{\sqrt{\frac{2MS_E}{n}}}$$



25

One technique for doing post hoc analysis Fisher's least significant difference method, the Fisher's least significant difference method compares all pairs of mean with the null hypothesis

$H_0: \mu_i = \mu_j$ for all $(i \neq j)$ using the t statistics this is nothing but your two sample t-test. So, $y_{i*} - y_{j*}$ divided by root of how we got this one generally we will get this one MSE divided by $n +$ MSE divided by n both are same we wrote 2 into MSE divided by n but here is the sample size is same.

(Refer Slide Time: 17:56)

Fisher's least significant difference (LSD) method

- Assuming a two-sided alternative hypothesis, the pair of means i and j would be declared significantly different if

$$\left| \bar{y}_{i*} - \bar{y}_{j*} \right| > LSD$$

where LSD, the least significant difference, is

$$LSD = t_{\alpha/2, a(n-1)} \sqrt{\frac{2MS_E}{n}}$$


26

Assuming a two-sided alternative hypothesis the pair of means i and j would be declared significantly different if the absolute value of the difference of their mean if it is greater than LSD then we will say that there is a significant difference between that 2 pair, so LSD if you bring the left hand side the previous formula will be $t_{\alpha/2, a(n-1)}$, a is the number of levels in this number of observations each treatment, $\sqrt{(2MSE \text{ divided by } n)}$ just I have readjusted that form.

(Refer Slide Time: 18:34)

Fisher's least significant difference (LSD) method

- If the sample sizes are different in each treatment, the LSD is defined as

$$LSD = t_{\alpha/2, N-a} \sqrt{MS_E \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$



27

So if the sample sizes are different in each treatment the LSD is defined as LSD equal to $t_{\alpha/2, N-a}$ root of MSE (1 divided by n_i + 1 divided by n_j) that means for each pair there will be a different LSD because the sample size are different that should be very careful on that one.

(Refer Slide Time: 19:00)

Problem : LSD method

- We will apply the Fisher LSD method to the hardwood concentration experiment. There are $a = 4$ means, $n = 6$, $MSE = 6.51$, and $t_{0.025, 20} = 2.086$. The treatment means are

$$\begin{aligned}\bar{y}_1 &= 10.00 \text{ psi} \\ \bar{y}_2 &= 15.67 \text{ psi} \\ \bar{y}_3 &= 17.00 \text{ psi} \\ \bar{y}_4 &= 21.17 \text{ psi}\end{aligned}$$



28

We will apply the fisher's LSD method to the hardwood concentration experiment there are 4 means n equal to 6 MS is 6.51 so in alpha equal to 5% for 0.025 and 20 degrees of freedom the t value is 2.086 this was the mean of different treatment.

(Refer Slide Time: 19:24)

Problem : LSD method

- The value of LSD is:

$$LSD = t_{0.025, 20} \sqrt{\frac{2MS_E}{n}} = 2.086 \sqrt{\frac{2(6.51)}{6}} = 3.07$$

- Therefore, any pair of treatment averages that differs by more than 3.07 implies that the corresponding pair of treatment means are different.



29

So, when you substitute here LSD value is 3.07 so we have to compare see that this one 1 and 2, 1 and 3, 1 and 4 - 1 3 and 2 and 4 therefore any pair of treatment averages that differs by more than 3.07 implies that that corresponding pair of treatment means are different. So, what you have to do next step or to take any two pair of the mean you have to find their absolute difference if the absolute difference is greater than 3.07 we can conclude that that two pairs means are different.

This was already we have got this ANOVA table now we will go for this LSD test import mat first we will find out the t value t values - 1 x scipy.stats.t.ppf of 0.025, 20 because why I am taking - 1 because whether it is the right side value the t value should be positive. So, n equal to 6 MSE is already we know that it is this value 6.50 so LSD is I am writing this formula t multiplied by math.sqrt(2*MSE/n), we are getting 3.07. So this value we got it already 3.07 this 3.07.

(Refer Slide Time: 20:49)

Problem : LSD method

- The comparisons among the observed treatment averages are as follows:

$$4 \text{ vs. } 1 = 21.17 - 10.00 = 11.17 > 3.07$$

$$4 \text{ vs. } 2 = 21.17 - 15.67 = 5.50 > 3.07$$

$$4 \text{ vs. } 3 = 21.17 - 17.00 = 4.17 > 3.07$$

$$3 \text{ vs. } 1 = 17.00 - 10.00 = 7.00 > 3.07$$

$$3 \text{ vs. } 2 = 17.00 - 15.67 = 1.33 < 3.07$$

$$2 \text{ vs. } 1 = 15.67 - 10.00 = 5.67 > 3.07$$



31

Now we are going to take all the pair's first you will take 4 versus 1, 4 versus 2, 4 versus 3 then 3 versus 1, 3 verses 2 and 2 versus 1, so the absolute difference is 11.17, 5.50 this is greater than so what we can conclude mu 4 not equal to mu 1, but here you see that if this is less than 3.07 so what we have to conclude is this mu 3 equal to mu 2 all other pairs are different.

(Refer Slide Time: 21:29)

Problem : LSD method

- In this problem we see that there are significant differences between all pairs of means except 2 and 3
- This implies that 10 and 15% hardwood concentration produce approximately the same tensile strength and that all other concentration levels tested produce different tensile strengths



32

In this problem we see that there are significant differences between all the pairs of mean except 2 and 3 this implies that 10 % and 15% hardwood concentration produce approximately the same tensile strength there is means are equal that two means are equal. So, what we are concluding that 10% and 15 % hardwood concentration produce approximately the same tensile strength and that all other concentration levels tested produce different tensile strength.

(Refer Slide Time: 11:01)

The Tukey-Kramer Test for Post Hoc analysis

- Tells which population means are significantly different
- Done after rejection of equal means in ANOVA
- Allows pair-wise comparisons
- Compare absolute mean differences with critical range

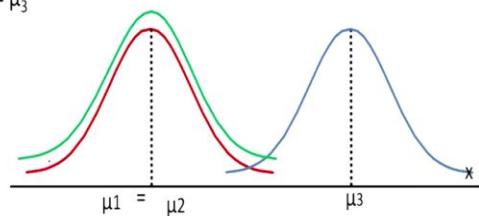
2

We will go for another post hoc analysis that is called Tukey Kramer test. Tukey Kramer test tells which population means are significantly different. It is then after rejection of equal means in ANOVA that means after rejecting our hypothesis it allows pairwise comparison all the means are compared as a pair. So, compare absolute mean differences with the critical range what will happen do in this test we will find out mean absolute difference so that difference is compared with the critical range that we got from the table called Tukey table.

(Refer Slide Time: 22:40)

The Tukey-Kramer Test for Post Hoc analysis

- Determine if there is any significant difference between the means
- Is $\mu_1 = \mu_2 \neq \mu_3$



3

So, Tukey Kramer test for post hoc analysis determine if there is any significant difference between the means so when we reject a null hypothesis this figure says that μ_1 equal to μ_2

but mu 3 is different. So, this which two pairs of means is equal that we can find with the help of this Tukey Kramer test.

(Refer Slide Time: 23:05)

Tukey-Kramer Critical Range

$$\text{Critical Range} = Q_U \sqrt{\frac{\text{MSW}}{2} \left(\frac{1}{n_j} + \frac{1}{n_{j'}} \right)}$$

where:

- Q_U = Value from Studentized Range
- Distribution with c and $n - c$ degrees of freedom for the desired level of α
- MSW = Mean Square Within
- n_j and $n_{j'}$ = Sample sizes from groups j and j'

4

So, here we have to find out the critical range the critical range is Q_U root of MSW or we can say MS within the column otherwise we can say MSE divided by 2 (1 divided by n_j plus 1 divided by $n_{j'}$),- n_j and $n_{j'}$ is 2 pairs which are comparing and corresponding sample size. Here the Q_U the value from studentized range that I will show you I have the table with me studentized range distribution with the c and $n - c$ degrees of freedom.

Here c is the number of columns n is the total number of sample size degrees of freedom for the desired level of alpha MSW is mean square within nothing but every a MSE n_j and $n_{j'}$ says or sample sizes from groups j and j' that means we are taking 2 pairs of population j and j' they are comparing the we are finding the absolute difference. If that absolute difference is greater than critical range we will say that that two pairs are different. If it is within the critical range we say that that 2 pairs means is same.

(Refer Slide Time: 14:20)

Problem: Tukey- Kramer test

- Tensile Strength of Paper (psi)

Hardwood Concentration (%)	Observations						Total	Avg
	1	2	3	4	5	6		
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96



5

This was the problem which have solved previously so for this problem we have remember we have rejected null hypothesis.

(Refer Slide Time: 24:30)

The Tukey-Kramer Procedure

- Compute absolute mean differences:

$$|\bar{x}_1 - \bar{x}_2| = |10.00 - 15.67| = 5.67$$

$$|\bar{x}_1 - \bar{x}_3| = |10.00 - 17.00| = 7$$

$$|\bar{x}_2 - \bar{x}_3| = |15.67 - 17.00| = 1.33$$

$$|\bar{x}_1 - \bar{x}_4| = |10.00 - 21.17| = 11.17$$

$$|\bar{x}_2 - \bar{x}_4| = |15.67 - 21.17| = 5.5$$

$$|\bar{x}_3 - \bar{x}_4| = |17.00 - 21.17| = 4.17$$



6

So first we are finding out reject a null hypothesis then we are going to do Tukey Kramer. So, we are going to compare mean of X 1 bar and X 2 bar, X 1 bar and X 3 bar, X 2 bar and X 3 bar X 1 X 4, X 2 X 4, X 3 and X 4. So, the absolute mean 4 X 1 bar - X 2 bar is 5.67 for second one is 7 for third one is 1.33, X 1 and X 4 is 11.14 and so on.

(Refer Slide Time: 15:04)

The Tukey-Kramer Procedure

2. Find the Q_U value from the table with $c = 4$ and $(n - c) = (24 - 4) = 20$ degrees of freedom for the desired level of α ($\alpha = .05$ used here):

$$Q_U = 3.96$$



Next we will find out the critical range find QU value from the table with see here number of treatment is 4, $n - c$ is 20 degrees of freedom for the desired allele of alpha equal to 5% this 3.96 is got from this table.

(Refer Slide Time: 251:22)

Error Term	2	3	4	5	6	7	8	9	10
5	3.64 5.70	4.60 6.98	5.22 7.80	5.67 8.42	6.03 8.91	6.33 9.32	6.58 9.67	6.80 9.97	6.99 10.24
6	3.46 5.24	4.34 6.33	4.90 7.03	5.30 7.56	5.63 7.97	5.90 8.32	6.12 8.61	6.32 8.87	6.49 9.10
7	3.34 4.95	4.16 5.92	4.68 6.54	5.09 7.01	5.36 7.37	5.61 7.68	5.82 7.94	6.00 8.17	6.16 8.37
8	3.26 4.75	4.04 5.64	4.53 6.20	4.89 6.62	5.17 6.96	5.40 7.24	5.60 7.47	5.77 7.68	5.92 7.86
9	3.20 4.60	3.95 5.43	4.41 5.96	4.76 6.35	5.02 6.66	5.24 6.91	5.43 7.13	5.59 7.33	5.74 7.49
10	3.15 4.48	3.88 5.27	4.33 5.77	4.65 6.14	4.91 6.43	5.12 6.67	5.30 6.87	5.46 7.05	5.60 7.21
11	3.11 4.39	3.82 5.15	4.26 5.62	4.57 5.97	4.82 6.25	5.03 6.48	5.20 6.67	5.35 6.84	5.49 6.99
12	3.08 4.32	3.77 5.05	4.20 5.50	4.51 5.84	4.75 6.10	4.95 6.32	5.12 6.51	5.27 6.67	5.39 6.81
13	3.06 4.26	3.73 4.96	4.15 5.40	4.46 5.73	4.69 5.98	4.88 6.19	5.05 6.37	5.19 6.53	5.32 6.67
14	3.03 4.21	3.70 4.89	4.11 5.32	4.41 5.63	4.64 5.88	4.83 6.08	4.99 6.26	5.13 6.41	5.25 6.54
15	3.01 4.17	3.67 4.84	4.06 5.25	4.37 5.56	4.59 5.80	4.78 5.99	4.94 6.16	5.06 6.31	5.20 6.44
16	3.00 4.13	3.65 4.79	4.05 5.19	4.33 5.45	4.56 5.72	4.74 5.92	4.90 6.08	5.03 6.22	5.15 6.35
17	2.98 4.10	3.63 4.74	4.02 5.14	4.30 5.42	4.52 5.66	4.70 5.85	4.86 6.01	4.99 6.15	5.11 6.27
18	2.97 4.07	3.61 4.70	4.00 5.09	4.28 5.38	4.49 5.60	4.67 5.79	4.82 5.94	4.96 6.08	5.07 6.20
19	2.96 4.05	3.59 4.67	3.98 5.23	4.25 5.55	4.47 5.73	4.65 5.89	4.79 6.02	4.92 6.14	5.04 6.14
20	2.95 4.02	3.58 4.64	3.96 5.00	4.23 5.29	4.45 5.51	4.62 5.68	4.77 5.84	4.90 6.09	5.01 6.09

Q table: The critical values
for q corresponding to
alpha = .05 (top) and
alpha = .01 (bottom)

When you look at this you see 4 is c , 20 is your $n - c$ so that corresponding value when alpha equal 0.05 and 3.96 okay the Q table the critical values for Q corresponding to alpha equal to 0.05 on top and 0.01 at the bottom. This is the ANOVA table this on our table says you see that MSE is 6.51 mean squared error, MS treatment is 127.6 because the value of 6.51 will use in the next slide.

(Refer Slide Time: 16:00)

The Tukey-Kramer Procedure

3. Compute Critical Range:

$$\text{Critical Range} = Q_U \sqrt{\frac{\text{MSW}}{2} \left(\frac{1}{n_j} + \frac{1}{n_j} \right)} = 3.96 \sqrt{\frac{6.51}{2} \left(\frac{1}{6} + \frac{1}{6} \right)} = 4.124$$

4. Compare:

$$|\bar{x}_1 - \bar{x}_2| = |10.00 - 15.67| = 5.67$$

$$|\bar{x}_1 - \bar{x}_3| = |10.00 - 17.00| = 7$$

$$|\bar{x}_2 - \bar{x}_3| = |15.67 - 17.00| = 1.33$$

$$|\bar{x}_1 - \bar{x}_4| = |10.00 - 21.17| = 11.17$$

$$|\bar{x}_2 - \bar{x}_4| = |15.67 - 21.17| = 5.5$$

$$|\bar{x}_3 - \bar{x}_4| = |17.00 - 21.17| = 4.17$$



So, third step is compute the critical range we will find the critical range QU which you got from the table root of MSW is 3.651 one which I shown in the previous table 5.6 divided by 2, 1 by 6 + 1 by 6 because same sample size, so that value is 4.124 then we will find the difference of two pair for example x_1 and x_2 the difference is 5.67 absolute difference is 5.67 but that is greater than 4.12 so we can say μ_1 not equal to μ_2 .

But look at this x_2 and x_3 this is less than your 4.124 so we will say μ_2 equal to μ_3 this is the observation which you got previously also so the mean of population 2 and population 3 is same, there is μ_2 and μ_3 same.

(Refer Slide Time: 27:00)

The Tukey-Kramer Procedure

5. Other then $|\bar{x}_2 - \bar{x}_3|$, all of the absolute mean differences are greater than critical range. Therefore there is significant difference between each pair of means, except 10% concentration and 15% concentration at the 5% level of significance.



Other than X 2 bar - X 3 bar when absolute value all of the absolute mean differences are greater than critical range therefore there is a significant difference between each pair of the means except 10 and 15% of concentration at 5% significance level. So, only these two concentrations there is no difference in tensile strength all other pairs are different.

(Refer Slide Time: 27:25)

```
In [53]: from statsmodels.stats.multicomp import pairwise_tukeyhsd
from statsmodels.stats.multicomp import MultiComparison
mc = MultiComparison(data_r1['value'], data_r1['treatments'])
mcresult = mc.tukeyhsd(0.05)
mcresult.summary()

Out[53]:
```

group1	group2	meandiff	lower	upper	reject
concentration10	concentration15	1.3333	-2.7894	5.4561	False
concentration10	concentration20	5.5	1.3773	9.6227	True
concentration10	concentration5	-5.6667	-9.7894	-1.5439	True
concentration15	concentration20	4.1667	0.0439	8.2894	True
concentration15	concentration5	-7.0	-11.1227	-2.8773	True
concentration20	concentration5	-11.1667	-15.2894	-7.0439	True

This we cannot solve with the help of Python so we import, from statsmodels.stats.multicomp import pairwise_tukeyhsd, honestly significant difference does that HSD from statsmodels.stats.multicomp import MultiComparison, mc = MultiComparison(data_r1['value'], data_r1['treatments']) is a value which you have already you remember we have done in the previous lecture data underscore onwards the treatment.

So mcresult equal to mc.tukeyhsd for 0.05 alpha mc result dot summary we will get this result. So, students what you have to do I have taken the screenshot of the output you have to enter this command into the Python command prompt then able to enter and verify the result. So, what is happening here is a group 1 group 2 you see 10%, 15% here false the rejection that means the rejection is false only that means these when mu equal to 10% of concentration and when the mu equal to 15% of concentration the means are equal all other pairs means are not equal.

(Refer Slide Time: 28:48)

Problem 2

- Following table shows observed tensile strength (lb/in square) of different clothes having different weight percentage of cotton.
- Check whether having different weight percentage of cotton, plays any role in tensile strength (lb/in square) of clothes.



We will do another problem the following table shows observed tensile strength found in lb/ in square of different clothes having different weight % of cotton check whether having different weight percentage of cotton plays any role in tensile strength what we are going to do in this problem whenever the percentage of cotton is added into thee into the clothes this tensile strength is increasing. We will see that is there any connection between percentage of cotton and the tensile strength of the clothes.

(Refer Slide Time: 29:23)

Problem 2

Weight Percentage of cotton	Observed tensile strength (lb/in square)					Total	Average
	1	2	3	4	5		
15	7	7	15	11	9	49	9.8
20	12	17	12	18	18	77	15.4
25	14	18	18	19	19	88	17.6
30	19	25	22	19	23	108	21.6
35	7	10	11	15	11	54	10.8
						Grand total=376	Grand mean= 15.004



So here the weight percentage is taken in drove 15%, 20%, 25, 30, 35 this was the 5 observationally is given total is given the grand total is 376 the grand mean is 15.07.

(Refer Slide Time: 29:39)

- $SSA = 5(9.8 - 15.04)^2 + 5(15.4 - 15.04)^2 + 5(17.6 - 15.04)^2 + 5(21.6 - 15.04)^2 + 5(10.8 - 15.04)^2 = 475.76$

$SST = 636.96$

$SSE = 636.96 - 475.76 = 161.20$

Sources of variation	Sum of squares	Degrees of freedom	Mean square	F-value
Cotton weight percentage	475.76	4	118.94	14.76
Error	161.20	20	8.06	
Total	636.96	24		



First we will find out SS treatment some books they follow in SSB that means between sum of square, SSA among sum of square some book write SS treatment, treatment sum of square. So, for treatment sum of square see in the row 1 there is a in the treatment 1 there are 5 element is there so 5 into this is the mean of first row. This is overall mean so 5 into 9.8 15.04 square plus this is their 5 element and 15.4 is the mean of the second row this is the overall mean Plus this is the mean of third row this is mean of third row mean of 4th row mean of your third row.

It is very getting SST treatment sum of square among the column that is otherwise SS treatment sum of sum of square is 475.76 similarly we have done previously with the help of our problems we can find out SSE we know this SST is 636.96 so if you want to know SSE 636.96 so we are getting 161.60 so this is ANOVA setup so this is sum of square of cotton weight percentage sum of square of error there is a degrees of freedom because there is a 5 rows, so $5 - 1 = 4$ rows so there are 25 elements $25 - 1 = 24$ so $24 - 4 = 20$ degrees of freedom.

When you divide by this 475.76 by 4 you are getting 118.94 when you divide 161.2 divided by 20 we are getting 8.06 so the F value is 4.76.

(Refer Slide Time: 31:19)

Problem 2

- When alpha =.05, $F_{0.05,4,20} = 2.87$
- Reject Ho

```
In [17]: scipy.stats.f.ppf(1-0.05, dfn=4, dfd=20)
Out[17]: 2.8660814020156584
```



When alpha equal to because we could 0.95 because 5% means 0.95 numerator degrees of freedom is 4 degrees of when we are getting so this value when alpha equal to 5% this is 2.8 okay but our calculated F value is 14.76 so 14.76 will be this side which is on the right hand side so we have to reject our null hypothesis. After rejecting null hypothesis we refer Q table.

(Refer Slide Time: 31:59)

Problem 2

$$T_a = q_a(c, n - c) \sqrt{\frac{M S_E}{n}}$$
$$\alpha = 0.05$$

$$q_{0.05}(5, 20) = 4.23$$
$$T_{0.05} = 4.23 \sqrt{\frac{8.06}{5}} = 5.37$$



So, the Q value which we got from the table is when alpha equal to 5% is in his 4.23 so when it is a 4.23 MSE is 8.06 we got 8.06 from this one this value MSE 8.06 this, this value is taken as 8.06 divided by n so it is a 5.37. So, this 5.37 you have to compare any pair of treatment averages that differ in absolute value by more than 5.37 would imply that corresponding pair of population means are significantly different.

(Refer Slide Time: 32:37)

Problem 2

$\bar{y}_1 - \bar{y}_2 = 9.8 - 15.4 = 5.6^*$	$\bar{y}_3 - \bar{y}_4 = 17.6 - 21.6 = 4$
$\bar{y}_1 - \bar{y}_3 = 9.8 - 17.6 = 7.8^*$	Starred values indicate pairs of means that are significantly different.
$\bar{y}_1 - \bar{y}_4 = 9.8 - 21.6 = 11.8^*$	
$\bar{y}_2 - \bar{y}_3 = 15.4 - 17.6 = 2.2$	$\bar{y}_3 - \bar{y}_5 = 17.6 - 10.8 = 6.8^*$
$\bar{y}_2 - \bar{y}_4 = 15.4 - 21.6 = 6.2^*$	$\bar{y}_4 - \bar{y}_5 = 21.6 - 10.8 = 10.8^*$
$\bar{y}_2 - \bar{y}_5 = 15.4 - 10.8 = 4.6$	



So, we have compared all comparison y_1 versus going to y_1 y_3 what is happening this is more than 5.37 this is this is here we can say μ_1 not equal to μ_2 here also μ_1 not equal to μ_3 but what is happening here it is less than, less than 5.37. So, here we can say μ_1 equal to μ_5 here also μ_2 equal to μ_3 here also μ_2 equal to μ_5 , here also μ_3 equal to μ_4 okay this is the Tukey Kramer test.

(Refer Slide Time: 33:22)

Jupyter code

```
In [2]: df3 = pd.read_excel('C:/Users/Soni/Documents/cotton weight.xlsx')
In [12]: data1 = pd.melt(df3.reset_index(), id_vars=['index'], value_vars=['cotwt.15','cotwt.20','cotwt.25','cotwt.30','cotwt.35'])
          data1.columns = ['id', 'treatments', 'value']
```



Then we will do with the help of Python I have imported the data calling it $df3$, `pd.read_excel` because I have saved in the excel format then I am using this melt command you know that previously in the two classes I have used how to use this melt what is application of this `pd.melt`

(df3_reset_index(), id_vars = ['index'], value_vars = ['cotton percentage 15, cotwt 20, cotwt 25 30, 35 so date data.Columns equal to ['id', 'treatment', 'values']].

(Refer Slide Time: 34:08)

```
In [16]: mc = MultiComparison(data1['value'], data1['treatments'])
mcresults = mc.tukeyhsd(0.05)
mcresults.summary()

Out[16]:
```

group1	group2	meandiff	lower	upper	reject
cotwt.15	cotwt.20	5.6	0.2266	10.9734	True
cotwt.15	cotwt.25	7.8	2.4266	13.1734	True
cotwt.15	cotwt.30	11.8	6.4266	17.1734	True
cotwt.15	cotwt.35	1.0	-4.3734	6.3734	False
cotwt.20	cotwt.25	2.2	-3.1734	7.5734	False
cotwt.20	cotwt.30	6.2	0.8266	11.5734	True
cotwt.20	cotwt.35	-4.6	-9.9734	0.7734	False
cotwt.25	cotwt.30	4.0	-1.3734	9.3734	False
cotwt.25	cotwt.35	-6.8	-12.1734	-1.4266	True
cotwt.30	cotwt.35	-10.8	-16.1734	-5.4266	True

So, mc is multi comparison data one value, data treatment one so, mcresult equal to mc.tukeyhsd for 0.05 when you mcresult.summary when you type this what is happening here this, this, this that means the corresponding means are equal. So, other places not equal, so we have got whatever we got that result we have checked with the help of Python also. Dear students in this lecture what you have seen we have solved one way ANOVA problem in that one way ANOVA we have rejected null hypothesis.

Once we reject null hypothesis we have to say which two pairs of means is equal or not equal for that we have gone for and another set of test called post hoc analysis there are two test was there one is the least significant difference method another rule is Tukey Kramer method and also this be solved another problem with the help of Python, thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 26
Randomize Block Design (RBD)

Dear students, the previous class we have seen one way Anova that is Completely Randomized Design, we call to CRD. In this class, we will see another technique called Randomized Block Design.

(Refer Slide Time: 00:42)

Learning Objectives

- Estimate variance components in an experiment involving random factors
- Understand the blocking principle and how it is used to isolate the effect of nuisance factors
- Design and conduct experiments involving the randomized complete block design

2

The class objectives are estimate the various components in experiment involving random factors, what will happen in Anova we are considering some factors. We are saying that the effect of the factor, what is the effect of the variance. But unknowingly, there is a possibility that some more variable may influence our response variable so that unknown variable and variance due to that unknown variables are going to remove it, then we are going to do the analysis.

Then, will see what is the effect of that one? Then, understand the blocking principle and how it is used to isolate the effect of nuisance factors. So, what you are doing here in Randomized block design. We are going to we are going to isolate the effect of nuisance factors then, design and conduct experiment involving Randomized Block design. A completely randomized design CRD

is useful when the experimental units are homogeneous. If the experiment units are heterogeneous blocking is often used to form homogeneous groups.

(Refer Slide Time: 01:50)

Why RBD?

- A problem can arise whenever differences due to extraneous factors (ones not considered in the experiment) cause the MSE term in this ratio to become large.
- In such cases, the F value in equation can become small, signaling no difference among treatment means when in fact such a difference exists.

$$F = \frac{\text{MSTR}}{\text{MSE}} \rightarrow$$

4

Why we have to go for RBD Randomized block design? A problem can arise whenever difference is due to extraneous factors that is, once not consider in the experiment cause the mean squared error term in this ratio to become large. What will happen? Due to that nuisance factor, the value of mean squared error will become very high. In such cases, f value in equation can become very small.

Signaling, no difference among treatment means when in fact such differences exist. So what will happen here in the MSE, there may be some error terms which are due to external factors. So we are going to find out how much error is due to external factor that we are going to remove it. Then, we are going to conduct the F Value.

(Refer Slide Time: 02:45)

Randomized block design

- Experimental studies in business often involve experimental units that are highly heterogeneous; as a result, randomized block designs are often employed.
- Blocking in experimental design is similar to stratification in sampling.

5

Experimental studies in business often involve experimental units that are highly heterogeneous as a result Randomized block designs are often employed. Blocking in experimental design is similar to certificate stratification in sample. In stratification in sampling what we are doing? We are if the samples are heterogeneous based on certain criteria we are grouping, we are stratifying that sample, so that each strata will have homogeneous sample.

(Refer Slide Time: 03:21)

Randomized block design

- Its purpose is to control some of the extraneous sources of variation by removing such variation from the MSE term.
- This design tends to provide a better estimate of the true error variance and leads to a more powerful hypothesis test in terms of the ability to detect differences among treatment means.

6

Here also, it is similar to stratification sampling. Its purpose is to control some of the external sources of variation by removing such variation from the MSE term. That is mean square error term. This design tends to provide a better estimate of the true error variance and leads to more powerful hypothesis test in terms of the ability to detect differences among treatment means.

(Refer Slide Time: 03:47)

Air Traffic Controller Stress Test

- A study measuring the fatigue and stress of air traffic controllers resulted in proposals for modification and redesign of the controller's work station.
- After consideration of several designs for the work station, three specific alternatives are selected as having the best potential for reducing controller stress
- The key question is: To what extent do the three alternatives differ in terms of their effect on controller stress?



7

We will take one sample example. This sample example is Air traffic controller stress test. Why this Air Traffic Controller is he has to schedule various aircraft what time it has to be landed, what time it has to take off. So, he is the person who has to allocate different slots for, for landing and takeoff. So this job is very stressful job. We will see one problem on this one. A study measuring the stress of Air traffic controller resulted in a proposal for modification and redesign of controller's workstation.

So, what they are planning? They are going to redesign, the work station because this sometime the workstation may influence, may affect the stress level. If it is the workstation is very narrow people are get stressed more ok. After consideration of several designs for the workstation, 3 specific alternatives are selected, as having the best potential of reducing controllers stress. They have identified the three alternatives.

The key question is to what extent do the three alternatives differ in terms of their effect on controller stress? So we are going to see to what extent they are different, different workstation design is going to affect the stress of the Air traffic controller.

(Refer Slide Time: 05:16)

Air Traffic Controller Stress Test

- In a completely randomized design, a random sample of controllers would be assigned to each work station alternative.
- However, controllers are believed to differ substantially in their ability to handle stressful situations.
- What is high stress to one controller might be only moderate or even low stress to another.
- Hence, when considering the within-group source of variation (MSE), we must realize that this variation includes both random error and error due to individual controller differences.
- In fact, managers expected controller variability to be a major contributor to the MSE term.



8

In a completely Randomized design a random sample of controllers would be assigned to each workstation alternative. Generally we will assign. However, Controllers are believed to differ substantially. It is not what we are assuming because the sample is not homogeneous because different controllers are affected by different level of workstation design. So, the controllers are believed to differ substantially in their ability to handle stressful situations.

What is high stress to one controller might be only moderate or even, low stress to another. So, what is happening? The sample is not homogeneous. Hence, when considering the within group sources of variation, that is MSE, let us call it means square error then, we must realize that this variation includes both random error and error due to individual control differences. In fact managers expected controller variability to be a major contribution to the MSE term.

(Refer Slide Time: 06:26)

A randomized block design for the air traffic controller stress test

Treatments

		System A	System B	System C
Blocks	Controller 1	15	15	18
	Controller 2	14	14	14
	Controller 3	10	11	15
	Controller 4	13	12	17
	Controller 5	16	13	16
	Controller 6	13	13	13



9

This is a set up. So what happened? There are three workstations design. We call it system, A system, B system, C. See this is a controller 1. So, in controller 1, when we put into system 1, the stress level is measured in terms of 15. So when controller 1, when he was subjected to work design 2 workstation design B he was expecting the stress of the distance is measured in terms of a questionnaire, so the 15 is the score, higher the score, higher the stress.

So controller 1, controller 2, controller 3, controller 4 controller, there are 6 controller. Since each controllers are different, if they are not homogeneous we are going to block it.

(Refer Slide Time: 07:08)

Solving this example using ANOVA in python

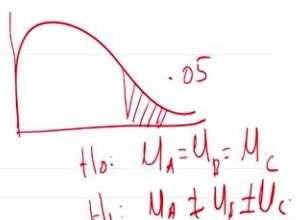
ANOVA

```
In [20]: data = pd.melt(df.reset_index(), id_vars=['index'], value_vars=['System A','System B','System C'])
data.columns = ['index', 'treatments', 'value']
```

```
In [21]: model = ols('value ~ C(treatments)', data=data).fit()
anova_table = sm.stats.anova_lm(model, typ=1)
anova_table
```

```
Out[21]:
      df  sum_sq  mean_sq    F   PR(>F)
C(treatments)  2.0  21.0  10.500000  3.214286  0.068903
Residual     15.0  49.0  3.266667    NaN    NaN
```

```
In [9]: # accept the null hypothesis
```



11

We solve these examples using Anova in Python. So, whether to import Pandas as pd, import numpy as np, import scipy, import statsmodels.api as sm, from statsmodels.formula.api, import ols. What I have done? The data which is in the table, which was in the previous slides, I have typed in an Excel. Then, I have imported that file name is RBD.xlsx so that I am going to save in the name of data frame. When I show the output this was system A, system B, system C.

So, I am using a melt command so data equal to pd.melt(df.reset_index(), id_vars = ['index'], value_vars = ['system A', 'system B', 'system C]), data.columns =[‘ index’, ‘ treatment’, ‘ value’]. This I also told you in the previous class. This melt command is used to bring all the values into column. One column is for treatment another column for values. So model equal to ols.

So, the value is the dependent variable tilde, See the treatments, data equal to that file name is data.fit. So anova_table equal to sm.stats.anova_lm, lm is a linear model. (model, typ =1). Remember this Type 1 because when whenever there is a two way anova you have to use type 2. So, Anova underscore table. What is happening? This error sum of square is 3.2. So, what is happening this value is more than 0.05.

So, we are accepting null hypothesis. What is the meaning of accepting null hypothesis? Here, so if it is a 0.05, so it is the P value. 06 we accepted null hypothesis. When I accept null hypothesis, what is the null hypothesis here? The level of stress is equal for different 3 workstation design. Suppose, H_0 equal to work station, This is stress, average stress level for workstation A, workstation design A, this is B, C. So $H_1: \mu_A \neq \mu_B \neq \mu_C$. So, at present what I am concluding I did not block it what time concluding? There is no connection between workstation design and the, and their average level of stress.

(Refer Slide Time: 10:02)

Summary of stress data for the air traffic controller stress test					
Treatments Blocks	System A	System B	System C	Block total	Block means
Controller 1	15	15	18	48	$\bar{x}_1 = 16$
Controller 2	14	14	14	42	$\bar{x}_2 = 14$
Controller 3	10	11	15	36	$\bar{x}_3 = 12$
Controller 4	13	12	17	42	$\bar{x}_4 = 14$
Controller 5	16	13	16	45	$\bar{x}_5 = 15$
Controller 6	13	13	13	39	$\bar{x}_6 = 13$
Column Total	81	78	93	252	$\bar{x} = 252/18 = 14$



12

Next, what I am going to do is I am going to do blocking going back. So there are 3.2 there is error. Actually 49, 49 divided by 15 so this is sum of square error is 49 mean sum of square is 3.2. So, this 3.26 due to blocking effect I am going to remove or subtract certain level of variance this. Then again, I am going to conduct, let us see what is happening. So this was the given data slide.

(Refer Slide Time: 10:33)

Summary of stress data for the air traffic controller stress test					
<ul style="list-style-type: none"> Treatment means 					
			$\bar{x}_1 = 81/6 = 13.5$		
			$\bar{x}_2 = 78/6 = 13$		
			$\bar{x}_3 = 93/6 = 15.5$		



13

So the treatment mean is there are three treatments that I am calling this system A system B system C. So, x_1 bar is 13.5, x_2 bar is 13, x_3 bar is 15.5 ok.

(Refer Slide Time: 10:52)

ANOVA TABLE FOR THE RANDOMIZED BLOCK DESIGN WITH k TREATMENTS AND b BLOCKS					
Sources of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	P- value
Treatments	SS Treatments	$k-1$	$MS_{Treatments} = \frac{SSTR}{k-1}$	$MS_{Treatments} / MSE$	
Blocks	SS block	$(b-1)$	$MS_{BL} = \frac{SSBL}{b-1}$		
Error	SSE	$(k-1)(b-1)$	$MSE = \frac{SSE}{(k-1)(b-1)}$		
Total	SST	$nr-1$			



14

We know this is our Anova setup. Look at the previous when there is the CRD in the completely Randomized design or one way Anova, there is no, there is no column blocking. Ok. There was only treatment and error. Now we are introducing the blocking so what is happening what are the degrees of freedom? Total number of element minus 1, degrees of freedom in treatment, there are k treatment $k - 1$ blocking, there are b blocking $b-1$.

So how to find out the $k-1$, $b-1$ is $n_T - 1 - (k-1) - (b-1)$, so we get $k-1$ and $b-1$. MSE treatment is SSE sum of square treatment divided by $k - 1$. MS mean square blocking equal to SSBL. Actually this data will not use that one. We will use only MS treatment by this MSE. SSE divided by $k-1$ into $b-1$. Actually this much portions, we will remove this will be subtracted. Ok.

(Refer Slide Time: 12:03)

RBD Problem

x_{ij} = value of the observation corresponding to treatment j in block i

\bar{x}_{ij} = sample mean of the j th treatment

\bar{x}_i = sample mean for the i th block

$\bar{\bar{x}}$ = overall sample mean

15

So x_{ij} is the value of the observation corresponding to the treatment j in the block i . \bar{x}_{ij} bar is the sample mean of j th treatment x_i . bar sample mean of i th block, $\bar{\bar{x}}$ double bar is overall sample mean.

(Refer Slide Time: 12:22)

RBD Problem

Step 1. Compute the total sum of squares (SST).

$$SST = \sum_{i=1}^b \sum_{j=1}^k (x_{ij} - \bar{\bar{x}})^2$$

Step 1. $SST = (15 - 14)^2 + (15 - 14)^2 + (18 - 14)^2 + \dots + (13 - 14)^2 = 70$

Step 2. Compute the sum of squares due to treatments (SSTR).

$$SSTR = b \sum_{j=1}^k (\bar{x}_{ij} - \bar{\bar{x}})^2$$

Step 2. $SSTR = 6[(13.5 - 14)^2 + (13.0 - 14)^2 + (15.5 - 14)^2] = 21$

16

What is the step 1? First, we will find out the SST that is a total sum of square. Total sum of squares summation i equal 1 to b , summation j equal 1 to j x_{ij} - individual element - overall main whole square. Ok. So in that way, we are getting $SST = 70$. Then compute the sum of square due to treatment so there are 3 treatments so, b is number of replication treatment 1. That is 6 column 1 in 13.5 - overall mean 14 whole square + 6 is common be brought 6 everything is brought in this side. So, $(13.0 - 14)^2 + (15.5 - 14)^2$ is 21.

(Refer Slide Time: 13:17)

RBD Problem

Step 3. Compute the sum of squares due to blocks (SSBL).

$$SSBL = k \sum_{i=1}^b (\bar{x}_{i\cdot} - \bar{\bar{x}})^2$$

$$\text{Step 3. } SSBL = 3[(16 - 14)^2 + (14 - 14)^2 + (12 - 14)^2 + (14 - 14)^2 + (15 - 14)^2 + (13 - 14)^2] = 30$$

Step 4. Compute the sum of squares due to error (SSE).

$$SSE = SST - SSTR - \text{SSBL}$$

$$\text{Step 4. } SSE = 70 - 21 - 30 = 19 //$$

17

So, 3rd step is compute the sum of square due to blocks. Due to blocks is there is a k treatment ok, so $\bar{x}_{i\cdot}$ minus $\bar{\bar{x}}$ so that row wise what is the mean? Everywhere there are 3 treatment 3 this one. So, 16 - 14 how we got the 16, going back to this 16, $(16 - 14)^2 + (14 - 14)^2$ square and so on. $(16 - 14)^2 + (14 - 14)^2 + (12 - 14)^2 + (14 - 14)^2 + (15 - 14)^2 + (13 - 14)^2$ is equal to 30.

This much variance is due to blocking, this much sum of square to compute the sum of square due to error term. We know that from SST you have to subtract treatment sum of square minus block in sum of square. That will give $SSE = 70 - 21 - 30$ is 19. So this 19 is the true SSE because this SSB amount which is due to extraneous variable that we are noise variable. So, error due to this, we are removing this.

(Refer Slide Time: 14:47)

ANOVA table for the air traffic controller stress test

Sources of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	P- value
Treatments	21	2	10.5	=5.53	0.024
Blocks	30	5	6.0		
Error	19	10	1.9		
Total	70	17			

$$F_{.025} = 5.46 \text{ and } F_{.01} = 7.56.$$

Reject the null hypothesis



18

While finding SSE now what we are getting. Yeah, the, whatever value which are given I kept it here. So it is a 10.5 divided by 1.9. Even though we find this one will not used for calculation. So 5.59 to the value of P value is 0.024 if Alpha equal to 5 percentage, we have to reject the null hypothesis. Previously what has happened? When we do without blocking, we are accepted null hypothesis and going back see, we accepted null hypothesis without blocking.

After blocking, our decision has completely changed. So what happened? We have rejected the null hypothesis.

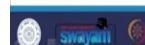
(Refer Slide Time: 15:34)

Solving RBD example using python

```
In [1]: import pandas as pd
import numpy as np
import scipy
import statsmodels.api as sm
from statsmodels.formula.api import ols
```

```
In [4]: df = pd.read_excel('RBD.xlsx')
df
```

	System A	System B	System C
0	15	15	18
1	14	14	14
2	10	11	15
3	13	12	17
4	16	13	16
5	13	13	13



19

We will do with the help of python import Pandas as pd, Import numpy as np, import scipy, import statsmodel.api as sm, from statsmodel.formula.api, import ols. ols is ordinary least square method because the regression and Anova is like two sides of the same coin. The sequence of learning Regression and Anova is first you have to learn Anova then you have to learn regression because there is a close relationship that I will see after this lecture is over after 2 lectures will go for Regression Analysis, I have imported.

(Refer Slide Time: 16:13)

```
In [20]: data = pd.melt(df.reset_index(), id_vars=['index'], value_vars=['System A','System B','System C'])
data.columns = ['blocks', 'treatments', 'value']

In [22]: model = ols('value ~ C(block)+ C(treatments)', data=data).fit()
anova_table = sm.stats.anova_lm(model, typ=1)
anova_table

Out[22]:
      df  sum_sq  mean_sq      F    PR(>F)
C(block)  5.0    30.0     6.0  3.157895  0.057399
C(treatments)  2.0    21.0    10.5  5.526316  0.024181
Residual  10.0    19.0     1.9    NaN      NaN

In [23]: # reject the null hypothesis
```

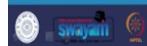
Ok There are 3 columns that I am using the melt Command so that I will bring all the values in two columns. One is for here there are 3 columns that are going to do the blocking. Blocks treatments and values so model equal to ols(‘value tilde C (block) + C(treatment)’) you see that now there is a blocking that I have included. Previously, there is no this term C x blocking so that I could data close bracket dot fit.

So, sm.stats.anova_lm (model, is a typ = 1), anova_table. I am getting you see this one here is this is 0.024 so it is less than 0.025 I am rejecting the null hypothesis.

(Refer Slide Time: 17:07)

Conclusion

- Finally, note that the ANOVA table shown in Table provides an F value to test for treatment effects but *not* for blocks.
- The reason is that the experiment was designed to test a single factor—work station design.
- The blocking based on individual stress differences was conducted to remove such variation from the MSE term.
- However, the study was not designed to test specifically for individual differences in stress.



21

So what we are concluding. Finally note that Anova table shown in the table provides f value test for treatment effect but not for the blocks. The reason is that experiment was designed to test a single factor workstation design. The blocking based on the individual stress, stress differences was conducted to remove such variation from the MSE term. However, the study was not designed to test specifically for individual differences in stress.

What is happening here is, the blocking exactly what you are doing? The error due to blocking is removed while finding the influence of workstation design on stress level.

(Refer Slide Time: 17:51)

Problem 2: RBD

- An experiment was performed to determine the effect of four different chemicals on the strength of a fabric.
- These chemicals are used as part of the permanent press finishing process.
- Five fabric samples were selected, and a randomized complete block design was run by testing each chemical type once in random order on each fabric sample.
- The data are shown in Table.
- We will test for differences in means using an ANOVA with alpha = 0.01.



22

We will go for one more problem will use this Randomized block design. We will go for one more problem. An experiment was performed to determine the effect of four different Chemicals on strength of your fabric. These Chemicals are used as a part of permanent press finishing process. 5 fabric samples were selected. And a Randomized complete block design was run by testing each chemical type once in a random order on each fabric sample.

The data is shown in the table in the next slide. We will test the difference in using Anova with Alpha equal to 1 percentage.

(Refer Slide Time: 18:29)

Problem 2: RBD							
Chemical Type	Fabric Sample					Treatment Totals y_i	Treatment Averages \bar{y}_i
	1	2	3	4	5		
1	1.3	1.6	0.5	1.2	1.1	5.7	1.14
2	2.2	2.4	0.4	2.0	1.8	8.8	1.76
3	1.8	1.7	0.6	1.5	1.3	6.9	1.38
4	3.9	4.4	2.0	4.1	3.4	17.8	3.56
Block totals y_{ij}	9.2	10.1	3.5	8.8	7.6	39.2($y_{..}$)	
Block averages \bar{y}_{ij}	2.30	2.53	0.88	2.20	1.90		1.96($\bar{y}_{..}$)



This was the table. What says this? Different chemical type is there, different fabric samples are there. The replication is five because the same after adding chemical Type 1 when we conduct the fabric strength, we have conducted 5 samples. This was the row mean, this was the row average.

(Refer Slide Time: 18:49)

Anova using jupyter

```

In [3]: df = pd.read_excel('rbd2.xlsx')
df
Out[3]:
   chem1  chem2  chem3  chem4
0     13     22     18     39
1     16     24     17     44
2     05     04     06     20
3     12     20     15     41
4     11     18     13     34

In [4]: data = pd.melt(df.reset_index(), id_vars=['index'], value_vars=['chem1','chem2','chem3','chem4'])
data.columns = ['index', 'treatments', 'value']

In [6]: model = ols('value ~ C(treatments)', data=data).fit()
aov_table = sm.stats.anova_lm(model, typ=1)
aov_table
Out[6]:
      df  sum_sq  mean_sq         F    PR(>F)
C(treatments)  3.0  10.044  6.01667  12.50569  0.000176
Residual      16.0   7.644  0.477759    NaN      NaN

```

24

What will you do? I have typed this data in excel, in excel file RBD2.xlsx. So this was the data so using melt coming and going to bring into the two variables. One is on value that is a response variable and other one is treatment. This you have to type as it is. That is the purpose of this pd.melt. So when I am running model = ols, a value is the dependent variable tilde, is the treatment is the independent variable.

Data is equal to data because this data is the way I have taken this after using melt command is the data. The file name is data so I am using data, data.fit and also another variable_table = sm.stats.anova_lm (model, is a typ = 1). So what is happening? You see that mean this is 0.4777. Here we did not do the blocking. We will do the blocking and what is happening to you? We are rejecting the null hypothesis because the probability is less than 0.01.

(Refer Slide Time: 20:00)

Problem 2: RBD

- The sums of squares for the analysis of variance are computed as follows:

$$\begin{aligned}
 SS_T &= \sum_{i=1}^4 \sum_{j=1}^5 y_{ij}^2 - \frac{\bar{y}_{..}^2}{ab} \\
 &= (1.3)^2 + (1.6)^2 + \dots + (3.4)^2 - \frac{(39.2)^2}{20} = 25.69 \\
 SS_{Treatments} &= \sum_{i=1}^4 \frac{\bar{y}_{i..}^2}{b} - \frac{\bar{y}_{..}^2}{ab} \\
 &= \frac{(5.7)^2 + (8.8)^2 + (6.9)^2 + (17.8)^2}{5} - \frac{(39.2)^2}{20} = 18.04
 \end{aligned}$$



25

So we are finding SST. SS treatment is see this formula is so comfortable for using calculator. So y_{ij} whole square minus $\bar{y}_{..}$ dot these notations, already I have explained. What is a ? a is the number of treatment b is the number of blocks? So this is SST is 25.69 is the total sum of square treatment sum of square is 18.04.

(Refer Slide Time: 20:27)

Problem 2: RBD

$$\begin{aligned}
 SS_{Blocks} &= \sum_{j=1}^5 \frac{\bar{y}_{.j}^2}{a} - \frac{\bar{y}_{..}^2}{ab} \\
 &= \frac{(9.2)^2 + (10.1)^2 + (3.5)^2 + (8.8)^2 + (7.6)^2}{4} - \frac{(39.2)^2}{20} = 6.69 \\
 SS_E &= SS_T - SS_{Blocks} - SS_{Treatments} \\
 &= 25.69 - 6.69 - 18.04 = 0.96
 \end{aligned}$$



26

SS block is $\bar{y}_{.j}$ whole square minus $\bar{y}_{..}$ double dot whole square divided by ab . So, this 6.69, is the error term, you see that of finding SST. Total sum of square minus sum of square due to blocking that I am subtracting that is due to treatment so I am getting 0.96. So this is a true error without having blocking effect.

(Refer Slide Time: 20:54)

Problem 2: RBD

- Analysis of Variance for the Randomized Complete Block Experiment

Sources of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	P- value
Chemical types (Treatments)	18.04	3	6.01	75.13	4.79 E-8
Fabric samples (Blocks)	6.69	4	1.67		
Error	0.96	12	0.08		
Total	25.69	19			

So what is happening? The mean square is here 0.08. So, you go back. What was the mean square without blocking? Yah, you see that it was without blocking it is 0.47 now it is 0.08. The mean square error is removed because we have removed the error due to blocking. So here the value of F also when you can compare it, it is significantly high. 75. 13 that is a more chances for rejection. Previously, what is the F value. I am going back. Here the values is 12. 12.58 now F value is 75.13. You are certainly you can say that you will reject your null hypothesis.

(Refer Slide Time: 21:50)

Conclusion

- The ANOVA is summarized in the previous table
- Since $f_0 = 75.13 > f_{0.01,3,12} = 5.95$ (the P-value is 4.79×10^{-8}), we conclude that there is a significant difference in the chemical types so far as their effect on strength is concerned.

Your Anova is summarised in the previous table. Since f equal to 75.13 which is greater than the table value that is a 5.95 which we got from the table, we have done Anova also. So that is the P

value is very low we conclude there is a significant difference in the chemical types so far as their effect of the strength is concerned.

(Refer Slide Time: 22:11)

Python code for problem 2

```
In [2]: import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

In [3]: df = pd.read_excel('RBD2.xlsx')

In [4]: df
```

Out[4]:

	chem1	chem2	chem3	chem4
0	1.3	2.2	1.8	3.9
1	1.6	2.4	1.7	4.4
2	0.5	0.4	0.6	2.0
3	1.2	2.0	1.5	4.1
4	1.1	1.8	1.3	3.4

Previously, we have done a traditional way. Now we will use Python for doing the blocking that is doing the Randomized block design. import pandas as pd, import statsmodels.api as sm, from statsmodels.formula.api import ols from statmodels.stats.anova import anova_lm. So, you save the file in the name df equal to pd.read_excel (). df This was the output.

(Refer Slide Time: 22:41)

Python code for problem 2

```
In [7]: data = pd.melt(df.reset_index(), id_vars=['index'], value_vars=['chem1','chem2','chem3','chem4'])
data.columns = ['Fabric samples', 'Chemical types', 'value']

Out[7]:
```

Fabric samples	Chemical types	value
0	chem1	1.3
1	chem1	1.6
2	chem1	0.5
3	chem1	1.2
4	chem1	1.1
5	chem2	2.2
6	chem2	2.4
7	chem2	0.4
8	chem2	2.0
9	chem2	1.8
10	chem3	1.8
11	chem3	1.7
12	chem3	0.6
13	chem3	1.5
14	chem3	1.3
15	chem4	3.9
16	chem4	4.4
17	chem4	2.0
18	chem4	4.1
..

Again, you see that we are using melt command after giving the melt command the data has become this format. So what is happening? Fabric samples 01234, 01234 see, these are 1 group.

This is another group. This is another group. This is another group, another group. So, this is chemical 1, chemical 2, chemical3, chemical4 the purpose of this pd.melt command is for this purpose.

Now there are three columns. One is the fabric sample. So the value, value is dependent variable chemical type treatment is independent variable. Fabric sample, that is, blocking variables.

(Refer Slide Time: 23:28)

The screenshot shows a Jupyter Notebook cell with the title "Python code for problem 2". The code is as follows:

```
In [11]: model = ols('value ~ C(Fabric) + C(Chemical)', data=data).fit()
anova_table = sm.stats.anova_lm(model, typ=1)
anova_table
```

The output is:

```
Out[11]:
```

	df	sum_sq	mean_sq	F	PR(>F)
C(Fabric)	4.0	6.693	1.673250	21.113565	2.31891e-05
C(Chemical)	3.0	18.044	6.014667	75.894849	4.51831e-08
Residual	12.0	0.951	0.079250	NaN	NaN

Now past this model is equal to ols ('value tilde C (fabric), fabric is, is a blocking effect plus the chemical that is the treatment effect, data equal to data.fit. When you run this we are getting see the f value is which we got traditionally manual method. We got this one so we see that P value. So what we have done, we taken one problem, we have solved without blocking what was the status. In this problem we are rejecting then we go for blocking.

After blocking also we are rejecting. But the when you look at the value of f that is significantly, it has increased. So, what will happen without blocking you may conclude on things? You may accept null hypothesis, because the error term is very bigger. After blocking the error term become very less than you may reverse the decision. We can reject the null hypothesis. That is application of this blocking.

Dear Students, in this lecture, what we have seen just I am summarizing. We have seen what is randomized block design? We have seen what is the need when will we go for a Randomized block design. Then you have taken your problem that problem was solved without blocking and seen what was the result then the same problem with blocking. Then you have seen how the result has changed. Even without blocking also we are used python code we saw what is the result?

Then, with blocking also we have used Python code. Then we have seen what was the result? In this, what we have done? We have taken two problems for both the problems we solved with blocking and without blocking. In the next class, we are going to another type of Anova that is a two way Anova. Thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 27
2 Way ANOVA

Dear students in the previous class we have seen randomized block design. In this class we will go to the next topic that is factorial experiments or 2 way ANOVA.

(Refer Slide Time: 00:37)

Learning objectives

- Design and conduct engineering experiments involving several factors using the factorial design approach
- Understand how the ANOVA is used to analyze the data from these experiments
- Know how to use the two-level series of factorial designs



2

The learning objectives are designed and conduct engineering experiments involving several factors using factorial design approach, understand how the ANOVA is used to analyze the data from these experiments and know how to use 2-level series of factorial design.

(Refer Slide Time: 00:58)

Factorial Experiment

- A **factorial experiment** is an experimental design that allows simultaneous conclusions about two or more factors.
- The term factorial is used because the experimental conditions include all possible combinations of the factors.
- The effect of a factor is defined as the change in response produced by a change in the level of the factor. It is called a main effect because it refers to the primary factors in the study
- For example, for a levels of factor A and b levels of factor B, the experiment will involve collecting data on ab treatment combinations.
- Factorial experiments are the only way to discover interactions between variables.



3

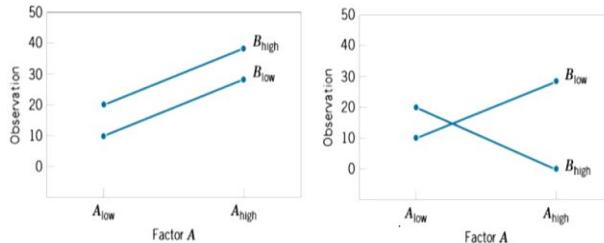
Let us see what is a factorial experiment a factorial experiment is an experimental design that allows simultaneous conclusions about 2 or more factors the previous 2 problems we did not see simultaneous effect of 2 variables at a time. The first problem when we whenever doing CRD completely randomized design we have taken only one independent variable the randomized to block design we have taken one independent variable and one blocking variable there was no interaction.

But in this lecture we will see that if there are 2 independent variable how there is a possibility of interactions we are going to see the effect of interaction also. The effect of factor is defined is the change in response produced by a change in the level of factor it is called main effect because it refers to the primary factors in the study. For example in levels of factor a and b levels of factor b small a is the level of factor a there may not be factor a there may be a levels small a levels may be low high there is a, 2 level.

For factor b also there may be 2 level low medium high for example 3 level also possible that is small b the experiment will involve collecting data on a b treatment combinations small a and small b. Factorial experiments are the only way to discover interaction between variables.

(Refer Slide Time: 02:27)

Factorial Experiment



Factorial Experiment, no interaction Factorial Experiment, with interaction

When you look at this diagram you see that there is a left side there is a factor that has 2 level low and high in observations you see the b and the factor they are also low a high, so these lines are parallel so these lines are parallel so that means that there is no interaction. When you look at the other side when the factor a goes low level to high level so what is happening there is a the crossing there is a interaction instead of a when a goes b is high when he is high then we also should be high.

But now b has become, it comes on the lower side so whenever there is a intersection that means that there is an interaction effect is there.

(Refer Slide Time: 03:15)

Two-factor Factorial Experiments

- The simplest type of factorial experiment involves only two factors, say, A and B.
- There are a levels of factor A and b levels of factor B.
- This two-factor factorial is shown in next table .
- The experiment has n replicates, and each replicate contains all ab treatment combinations.

The simplest type of factorial experiment involves only 2 factors say A and B there are a levels small a, a levels of factor A and b levels of factor B this 2 level factorial shown in the next table the experiment has n replicates and each replicate contains all a b treatment combinations.

(Refer Slide Time: 03:38)

Two-factor Factorial Experiments						
Data Arrangement for a Two-Factor Factorial Design						
		Factor B			Totals	Averages
Factor A	1	$y_{111}, y_{112},$ \dots, y_{11n}	$y_{121}, y_{122},$ \dots, y_{12n}	$y_{1b1}, y_{1b2},$ \dots, y_{1bn}	$y_{1..}$	$\bar{y}_{1..}$
	2	$y_{211}, y_{212},$ \dots, y_{21n}	$y_{221}, y_{222},$ \dots, y_{22n}	$y_{2b1}, y_{2b2},$ \dots, y_{2bn}	$y_{2..}$	$\bar{y}_{2..}$
	:					
	a	$y_{a11}, y_{a12},$ \dots, y_{a1n}	$y_{a21}, y_{a22},$ \dots, y_{a2n}	$y_{ab1}, y_{ab2},$ \dots, y_{abn}	$y_{a..}$	$\bar{y}_{a..}$
	Totals	$y_{..1}$	$y_{..2}$	$y_{..b}$	$y_{...}$	$\bar{y}_{...}$
Averages		$\bar{y}_{1..}$	$\bar{y}_{2..}$	$\bar{y}_{..b}$		

Look at this there is a factor, factor here there is a, a level is there factor b there are b level is there so maybe observations will be there the observation in a jth cell for the kth replicate is denoted by y_{ijk} in performing the experiment the a b and observations would be run in the random order. Thus like a single factor experiment the 2 factor factorial is a completely randomized design this is also kind of you CRD.

(Refer Slide Time: 04:10)

Example						
<ul style="list-style-type: none"> As an illustration of a two-factor factorial experiment, we will consider a study involving the Common Admission test (CAT), a standardized test used by graduate schools of business to evaluate an applicant's ability to pursue a graduate program in that field. Scores on the CAT range from 200 to 800, with higher scores implying higher aptitude. 						

We will take an example with the help of example we will see how to do the 2 way ANOVA. As an illustration of your 2 factorial experiments we will consider a study involving Common Admission Test for example in MBA suppose we want to get admission MBA you have to go this Common Admission Test. A standardized test used by Graduate School of Business to evaluate the applicants ability to pursue a graduate program in the field.

Scores on the CAT range from 200 to 800 in India it is seen it is expressed in terms of percentile assume that the range is 200 to 800. So, that means the minimum qualify marks for CAT is a in terms of absolute term say 200 not in terms of percentile see but the higher scores imply higher aptitude.

(Refer Slide Time: 05:04)

The slide has a dark blue header and footer. The title 'Three CAT preparation programs.' is in blue. The list below consists of a bullet point followed by three numbered options.

- In an attempt to improve students' performance on the CAT, a major university is considering offering the following three CAT preparation programs.
 1. A three-hour review session covering the types of questions generally asked on the CAT.
 2. A one-day program covering relevant exam material, along with the taking and grading of a sample exam.
 3. An intensive 10-week course involving the identification of each student's weaknesses and the setting up of individualized programs for improvement.

There are 3 CAT preparation programs in an attempt to improve students performance on the CAT a major university is considering offering the following 3 CAT preparation programs, there are 3 CAT population program the first program is 3 hour review, the second is one-day program the third one is intense you 10 weeks course involving. There are 3 type of coaching technique one is a 3-hour review session covering the types question generally asked in the CAT.

One day program covering relevant exam material along with; the taking and grading of sample exam. And intensive 10-week course involving the identification of each student's weaknesses and setting of individualized programs for improvement.

(Refer Slide Time: 05:04)

Factor - 1 , 3 treatment

- One factor in this study is the CAT preparation program, which has three treatments:
 - Three-hour review,
 - One-day program, and
 - 10-week course.
- Before selecting the preparation program to adopt, further study will be conducted to determine how the proposed programs affect CAT scores.

10

One factor is in this study is the CAT preparation program which has 3 treatment we are calling it as a treatment 3 hours review, one day program, 10 week course. There are 3 treatment before selecting the preparation programs to adopt further study will be conducted to determine how this proposed program effect to the CAT score. So, we are going to see there are 3 way of learning are preparing for the CAT examinations we are going to see the effect of these learning methods and how it is going to affect the performance in the CAT examination that is a CAT scores.

(Refer Slide Time: 06:34)

Factor 2 : 3 Treatment

- The CAT is usually taken by students from three colleges:
 - the College of Business,
 - the College of Engineering, and
 - the College of Arts and Sciences.
- Therefore, a second factor of interest in the experiment is whether a student's undergraduate college affects the CAT score.
- This second factor, undergraduate college, also has three treatments:
 - Business,
 - Engineering, and
 - Arts and sciences.

11

Factor 2 also there are 3 treatment CAT is usually taken by students from 3 colleges the college

those have undergraduate business school, the college who come from engineering backgrounds, the College of Arts and Sciences. Therefore a second factor of interest in the experiment is whether students undergraduate college effect to the CAT score. Now for example in many IITs the arts and science students are not allowed to take MBA examinations but I would prefer at present many IITs they are allowing even arts students also to get admitted into MBA program but they can take the CAT examinations.

Therefore we look at the will continue this problem therefore we a second factor of interest in the experiment is whether the students undergraduate college affect to the CAT score. The second factor undergraduate college also has a 3 treatment business a student may have Business Studies background or engineering background or art and science background. So, what we are going to study we are going to learn from this whether their undergraduate college or their background will affect their performance in the CAT's score.

Maybe sometimes the engineering students do better in CAT examinations, sometimes the background from the BBA students in the business background may do the CAT examination better, sometimes the arts and science students there may be a possibility because they may not come across many quantitative subjects in their undergraduate they there is a perception that they may not do well in the examination that we will see in this exam in this problem.

(Refer Slide Time: 08:21)

Nine Treatment Combinations for The Two-factor CAT Experiment

Factor A: Preparation Program	Factor B: College		
	Business	Engineering	Arts and sciences
Three-hour review	1	2	3
One-day program	4	5	6
10-Week course	7	8	9

12

So, what was done there are it is written in the table format in row it is taken the preparation program in factor in the column we take in college. So, what this table represents 9 treatment combinations for 2 factor CAT experiment, one factor is preparation program another factor is college they belongs to. So, you see that a person may be business background he may take 3 hours review he may be engineering background he may take 3 hours review he may be art and science background 3. So, there are 9 combinations are possible.

(Refer Slide Time: 09:04)

Replication

- In experimental design terminology, the sample size of two for each treatment combination indicates that we have two **replications**.

13

What is the replication an experimental design terminology the sample size for 2, for each treatment combination indicates that we have 2 replicates, if there are saying 2 for example here this 580 is the replication there are 2 students. So, what do you done a person who belongs to

business background when he undertake 3 hours review of taking coaching method, so what was their marks. So, we have subjected for 2 students we are subjected to 2 students.

Similarly 2 students are taking those who are engineering background and 3 our reviews, so how many 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 there are 18 observations.

(Refer Slide Time: 09:55)

The analysis of variance computations answers the following questions.

- **Main effect (factor A):** Do the preparation programs differ in terms of effect on CAT scores?
- **Main effect (factor B):** Do the undergraduate colleges differ in terms of effect on CAT scores?
- **Interaction effect (factors A and B):** Do students in some colleges do better on one type of preparation program whereas others do better on a different type of preparation program?

15

So, what is the analysis of variance computations answers the following questions what we are going to do that was the very important this lecture. So, what is the main effect factor a do the preparation programs differs in terms of effect on CAT's score. So, what we are going to see whether the different preparation programs affect their performance in the CAT course. Main effect B there is a factor B do the undergraduate college differs in terms of effect on CAT's course whether their undergraduate background is going to affect their performance in CAT's course are not.

Then interaction effect that is a factor A and factors B do students in some colleges do better than one type of preparation program whereas others do better on a different type of preparation programs that we are going to see interactions.

(Refer Slide Time: 10:54)

Interaction

- The term **interaction** refers to a new effect that we can now study because we used a factorial experiment.
- If the interaction effect has a significant impact on the CAT scores, we can conclude that the effect of the type of preparation program depends on the undergraduate college.

16

The term interaction refers to a new effect that we can now study because we used a factorial experiment if the interaction effect has significant impact on the CAT's course we can conclude that the effect of the type of preparation programs depends on the undergraduate college that is the learning. If the interaction is significant we can conclude that the type of preparation programs depending upon the, their undergraduate college background.

(Refer Slide Time: 11:29)

ANOVA Table for the Two-factor Factorial Experiment with r Replications

Sources of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	P- value
Factor A	SSA	(a-1)	SSA/a-1	MSA / MSE	
Factor B	SSB	(b-1)	SSB/b-1	MSB / MSE	
Interaction	SSAB	(a-1)(b-1)	MSAB = SSAB/(a-1)(b-1)	MSAB / MSE	
Error	SSE	ab(r-1)	MSE= SSE/(ab)(r-1)		
Total	SST	nT-1			

17

This is a two factor factorial experiment setup so Factor A is here SSEA sum of square for factor A, SSB sum of square for factor B SSE is sum of square for error so when you subtract it this is AB for interactions. So, how to know the how to write the degrees of freedom so n T - 1 that is a

degrees of freedom for total number of 'a' levels in 'a' ($a - 1$) degrees of freedom for factor b - 1 for factor B when you multiply $a - 1$ into $b - 1$ degrees of freedom for interaction ok.

Then you see that this is a mean square for factor a mean square factor B this is mean square for interaction this is a mean square error. So, what we are going to see we are going to see effect of factor A and effect of factor B and effect of interactions. If you want to know the effect of factor a we have to write mean sum of square for factor A divided by MSE if you want to know the effect of B means sum of square for factor B divided by MSE.

If we want to know the effect of interaction means sum of square MSAB divided by MSE you see that in the denominator always there is a error term. Many students will do mistakes there because MSA divided by MSE the denominator always there should be a error term.

(Refer Slide Time: 12:49)

The screenshot shows a presentation slide with a blue header and footer. The main title is 'Abbreviation' in blue. Below it, four definitions are listed:

- a = number of levels of factor A
- b = number of levels of factor B
- r = number of replications
- n_T = total number of observations taken in the experiment; $n_T = abr$

The footer contains three icons: a person, a book, and a gear, followed by the text 'Shajani'.

So, a represents number of levels of factor A, b represents number of levels of factor B, r represents number of replications n_T represents total number of observations taken in the experiment. so, n_T equal to abr because number of levels in a number of levels b material by number of applications.

(Refer Slide Time: 13:11)

ANOVA Procedure

- The ANOVA procedure for the two-factor factorial experiment requires us to partition the sum of squares total (SST) into four groups:
 - sum of squares for factor A (SSA),
 - sum of squares for factor B (SSB),
 - sum of squares for interaction (SSAB), and
 - sum of squares due to error (SSE).
- The formula for this partitioning follows.

$$SST = SSA + SSB + SSAB + SSE$$

19

The ANOVA procedure for 2 factor factorial experiment requires us to partition of sum of squares into 4 groups, sum of square so SST we are partitioning into sum of square due to factor A, sum of square for factor B sum of square for interaction ,of interaction and sum of squares due to the error term so, the formula for this partition follows SST equal to factor A sum of square plus factor B sum of square plus factor AB sum of square that is interaction sum of square plus error sum of square.

(Refer Slide Time: 13:48)

Computations and Conclusions

x_{ijk} = observation corresponding to the k th replicate taken from treatment i of factor A and treatment j of factor B

\bar{x}_i = sample mean for the observations in treatment i (factor A)

\bar{x}_j = sample mean for the observations in treatment j (factor B)

\bar{x}_{ij} = sample mean for the observations corresponding to the combination of treatment i (factor A) and treatment j (factor B)

\bar{x} = overall sample mean of all n_T observations

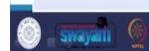
20

The notations are X_{ijk} observations corresponding to the k th replicate taken from the treatment i of factor and treatment j of factor b, X_i . bar represents sample mean for observation treatment i X_j bar represents sample mean for observation in treatment j factor B, X_{ij} bar represents sample

mean for the observations corresponding to the combination of the treatment i factor A and treatment j factor B X double bar is overall sample mean of all nT observations.

(Refer Slide Time: 14:33)

CAT Summary Data for The Two-factor Experiment					
Factor A: Preparation Program	Factor B: College			Row totals	
	Business	Engineering	Arts and sciences		
Three-hour review	500 $\bar{x}_{11} = 540$ 580	540 $\bar{x}_{12} = 500$ 460	480 $\bar{x}_{13} = 440$ 400	2960	
One-day program	460 $\bar{x}_{21} = 500$ 540	560 $\bar{x}_{22} = 590$ 620	420 $\bar{x}_{23} = 450$ 480	3080	
10-Week course	560 $\bar{x}_{31} = 580$ 600	600 $\bar{x}_{32} = 590$ 580	480 $\bar{x}_{33} = 445$ 410	3230	
Column totals	3240	3360	2670	Overall total= 9270	$\bar{x} = 515$



21

So, the first step is individually for each cell we have to find out the mean so here this location is X 1 1 bar mean is 540, X 1 2 bar mean is 500, X 1 3 mean is 500 the road total is 2960. So, the 2 1 mean is 500, 2 2 mean is 590, 2 3 mean is 550. So, third one X 3 1 bar equal to 580 X 3 2 = 590 X 3 3 = 445 the overall sum is 9270 the overall mean is 515.

(Refer Slide Time: 15:12)

CAT Summary Data for The Two-factor Experiment					
<ul style="list-style-type: none"> Factor A means $\bar{x}_{1.} = 493.33$ $\bar{x}_{2.} = 513.33$ $\bar{x}_{3.} = 538.33$ 					
<ul style="list-style-type: none"> Factor B means $\bar{x}_{.1} = 540$ $\bar{x}_{.2} = 560$ $\bar{x}_{.3} = 445$ 					
<p style="text-align: right;">22</p>					

Now the factor A means for row 1, X 1 bar equal to 493.33 X 2. bar is 513.33 X 3. bar equal to 538.33 for factor B means X you see that if it is a B only X. 1 bar is 540 X . 2 bar is 560 X. 3 is 445.

(Refer Slide Time: 15:45)

CAT Example:

Step 1. Compute the total sum of squares.

$$SST = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (x_{ijk} - \bar{x})^2$$

$$\text{Step 1. } SST = (500 - 515)^2 + (580 - 515)^2 + (540 - 515)^2 + \dots + (410 - 515)^2 = 82,450$$

23

First step is to find out total sum of square we know that total sum of square is $\sum_{ijk} x_{ijk}$ each element minus the overall mean whole square. So, what will happen we have to do for all the observations so that is coming 82,450. Step 2 compute the sum of square for factor A, so you write if you are writing you look at this if you are writing SSA so b into r, summation of \bar{x}_i dot bar that is we can say it is a row mean row mean minus overall mean whole square.

So SSA is 3 b equal to 3 replicas 2 because 2 data set so 493.33 is row 1 mean minus overall mean whole square, plus 513.33 is row 2 mean 515 whole square plus 538.33 minus overall mean whole square that is 60100.

(Refer Slide Time: 16:56)

CAT Example:

Step 3. Compute the sum of squares for factor B.

$$SSB = ar \sum_{j=1}^b (\bar{x}_j - \bar{x})^2$$

Step 3. $SSB = (3)(2)[(540 - 515)^2 + (560 - 515)^2 + (445 - 515)^2] = 45,300$



25

Now compute the sum of square of factor B whenever you write SSB see that a will come here ar summation j equal to 1 to b is nothing but your column mean, so column mean is 540 I am going back how we got the 540 and going back so 540, 580 this 580 540 560 445. Now let us go back next we are going to compute the sum of square for factor B so SSB equal to a.r summation j equal to 1 to be X dot J bar - X double bar whole square.

So there are a is 3 to 2 replications 540 is your column mean and going back how we got 540 560 445 and go back this is 540 560 445 that why got this one, that is why we got this one, 540 560 so SSB is 45300. Next we will go for SSAB compute the sum of square of interaction see here are in 2 there are 2 summations i equal to 1 to a j equal to 1 to b \bar{X}_{ij} minus that means in the cell what was the mean minus row mean minus column mean plus overall mean whole square.

So, \bar{X}_{aj} is in that cell mean is 540 row mean is 493.33 minus column mean is 540 plus overall mean is 515 whole square. So, SSAB when you continue this we are getting 11200.

(Refer Slide Time: 18:48)

CAT Example:

Step 5. Compute the sum of squares due to error.

$$SSE = SST - SSA - SSB - SSAB$$

Step 5. $SSE = 82,450 - 6100 - 45,300 - 11,200 = 19,850$

27

Then we will find out SSE, SSE is total if you subtract from the SST so total sum of square minus sum of square due to factor minus sum of square due to factor B minus sum of square due to A B so that SSE is 19,850.

(Refer Slide Time: 19:05)

ANOVA Table for the CAT two-factor design

Sources of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	P- value
Factor A	6100	2	3050	1.38	0.299
Factor B	45300	2	22650	10.27	0.005
Interaction	11200	4	2800	1.27	0.350
Error	19850	9	2206		
Total	82450	17			

28

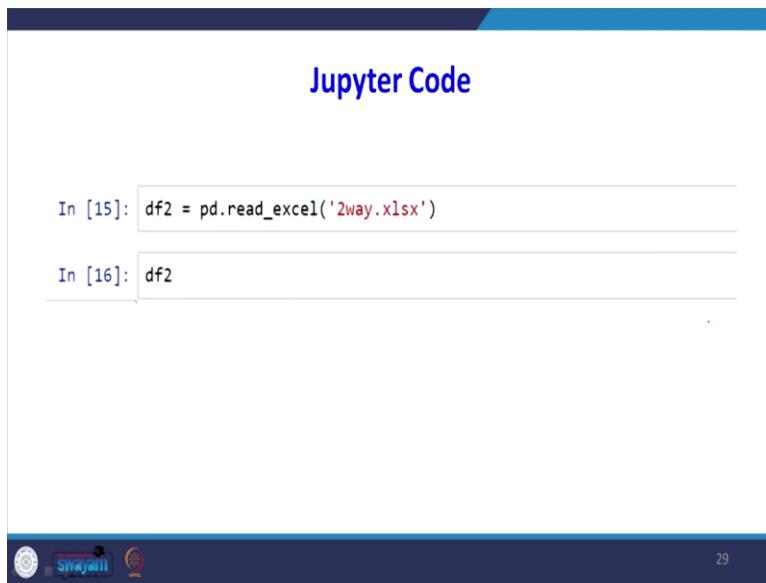
So, I have filled that value for factor a sum of 6100, 45300 interaction 11200 error 19850 so total sum of square is 82450, nothing but this 82450, 2450 is splitted into 4 parts one is due to factor A due to interaction in factor B and in due to interaction due to error. So, the degrees of freedom is there are 18 data set is there, so 18 - 1 is 17 there are 3 factors so 3 - 1 is 2 there are for factor A , for factor B there are 3 treatments so 3 – 2, 2 so we are getting a interactions it is number of levels in factor A that - 1 multiplied by number of levels in factor B that – 1.

So how we got this one go back you see that how we are getting the degrees of freedom for degrees of freedom for interaction right $a - 1$, a is number of level in factor a number of levels in factor that $- 1$ multiplied by number of levels in factor B that $- 1$. So, that value is this one you get 2 multiplied by 2 is 4. So, how we got this mean square 3050, when you divide 6100 by 2 3050 when you divide 43200 by 2, 22650, 11200 divided by 4 for 2,800.

So 19850 divided by nine this one so how we got F value is when you divide 3050 by 2206 you will get to 1.38 when you divide 22000 divided by 650 divided by 22650 get 10.27 when you divide 2800 by 2206 is at 1.27 this is a corresponding p-value. So, what does happy here, here we are accepting null hypothesis when you accept a null hypothesis there is no effect of factor A.

Here interaction we are accepting null hypothesis there is no effect of interaction but there is an effect of factor B because it is less than point 0.05.

(Refer Slide Time: 21:46)



Jupyter Code

```
In [15]: df2 = pd.read_excel('2way.xlsx')
In [16]: df2
```

The data whichever is given there I have entered into the in Excel in excel file, so I am reading df 2 equal to pd.read_excel.

(Refer Slide Time: 22:02)

Jupyter code		
Out[16]:		
Value	prep_pro	college
0	500	three_hr
1	580	three_hr
2	540	three_hr
3	460	three_hr
4	480	three_hr
5	400	three_hr
6	460	One-day
7	540	One-day
8	560	One-day
9	620	One-day
10	420	One-day
11	480	One-day
12	560	10-Week
13	600	10-Week
14	600	10-Week
15	580	10-Week
16	480	10-Week
17	410	10-Week

30

So, when I say df2 to see the data is in this permit value is in the first column there is 18 including 0 there are 18 values, see the preparation program 3 hours up to this there are 3 hours, first is 6 data set this is one day this is 10 week, this is those who are belongs to business background, those belongs to engineering background, those who belongs to art and science background. Again this is for whenever one day preparation program who belongs to business background, engineering background art science background when they go for 10 weeks intensive training program there also business background , engineering background and art and science background.

(Refer Slide Time: 22:55)

Jupyter Code					
In [20]:					
formula = 'Value ~C(college)+C(prep_pro)+C(college):C(prep_pro)'					
model = ols(formula, df2).fit()					
aov_table = anova_lm(model, typ=2)					
<pre>print(aov_table)</pre>					
C(college)	sum_sq	df	F	PR(>F)	
45300.0	2.0	10.269521	0.004757		
C(prep_pro)	6100.0	2.0	1.382872	0.299436	
C(college):C(prep_pro)	11200.0	4.0	1.269521	0.350328	
Residual	19850.0	9.0	NaN	NaN	

31

Here formula equal to ‘value tilde C(college) plus C(preparation program) plus C(college) colon C(preparation program)’. This represents for interaction, so model equal to ols (formula you can write it directly otherwise you can specify separately ols(formula ,df 2).fit so analysis of variance underscore table equal to anova_lm (model , see type 2, when you write typ = 2 it is for 2-way anova.

So, when we are writing for one-way anova here we write in typ = 1 we got one way ANOVA, so when you print on our table so we are getting this one. So, what is the meaning though the what we do it manually what you do in Python is same here what is happening this we accept a null hypothesis this we accept null hypothesis this we reject a null hypothesis. So, there is no interaction there is no effect of preparation program but there is an effect of the college background they belongs to.

They may be belongs to be engineering background they may be belongs to business background or art and science background because what we are concluding from here is that if there belongs to particular background there is courses their performance in CAT's score is otherwise we can say the college backgrounds affect their performance in the CAT's score it may be those who are belongs to engineering they can perform better or those who belongs to Arts and Science they may not perform better.

So what we are concluding here is the college they belongs to is an important variable on their performance in the CAT examination. We got the ANOVA table 2 way ANOVA table when we look at that see the preparation program is not a significant variable the interaction between college and the preparation program also not significant factor here but only the college there belongs to is an significant factor that means there are 3 possibility their college background may be one is Arts and Science second one is a business third one is engineering background. So that factor will affect their performance in the CAT's score.

Dear students in this class we have studied what is a 2 way ANOVA then we have taken one problem we have traditionally we have solved that 2 way ANOVA. Then I have explained the theoretical background behind this 2 way ANOVA then the same problem we have solved with

the help of Python. Then we have interpreted the result. The next class will go to another topic that is a regression analysis because this analysis of variance and regression analysis are it is like a 2 side of the same coin.

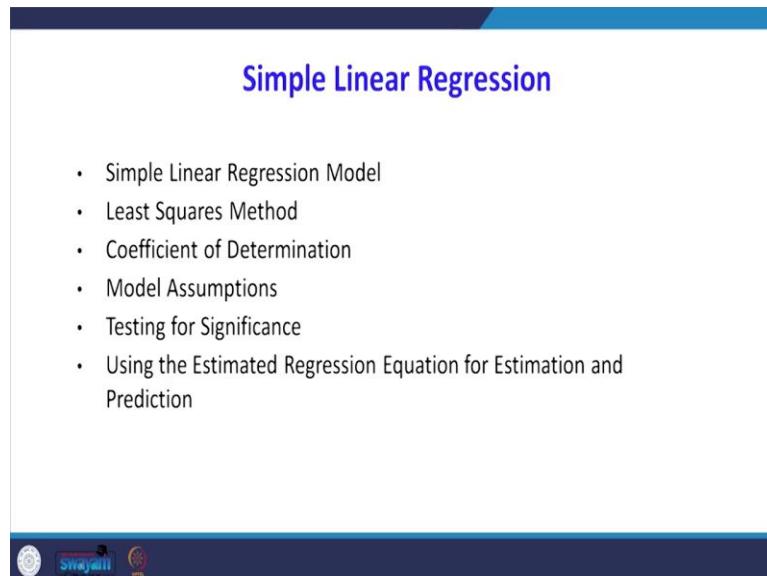
Even ANOVA can be solved with the help of regression analysis even a regression problem can be solved with the help of ANOVA. So, the next class I will meet you with another new topic called regression techniques, thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 28
Linear Regression - I

Dear students in this class we will go to the new topic called regression analysis, the class objectives of this lecturers is;

(Refer Slide Time: 00:34)



Simple Linear Regression

- Simple Linear Regression Model
- Least Squares Method
- Coefficient of Determination
- Model Assumptions
- Testing for Significance
- Using the Estimated Regression Equation for Estimation and Prediction



We will study simple linear regression model when you say simple linear regression model only one independent variable will be considered then will see what is the least square method. That is the principle behind this regression model. Then I will see what is coefficient of determination goodness of regression model explained, generally with the help of this coefficient of determination called R square we will see in detail later.

What are the model assumptions then we can test for significance, even hypothesis testing also can be done with the help of our regression analysis, then using the estimated regression equation for estimation and use the prediction also.

(Refer Slide Time: 01:19)

Empirical Models

- Many problems in engineering and science involve exploring the relationships between two or more variables
- Regression analysis is a statistical technique that is very useful for these types of problems
- This model can also be used for process optimization, such as finding the level of temperature that maximizes yield, or for process control purposes



3

Many problems in Engineering and Science involve exploring the relationship between 2 or more variables. So far what you have seen the same variable we have compared with the, we have taken some sample with help of sample we are predicted the population parameter the same variable sometime we are compared the mean sometime we compared variance but this lecture they are going to take 2 different variables.

Regression analysis is a statistical technique that is very useful for these type of problems where the cause and effect has to be measured. This model can also be used for process Optimisation such as finding the level of temperature that maximizes yield or process control purposes. There are many independent variable we can say which independent variable is more important variable that affect our dependent variable.

(Refer Slide Time: 02:12)

Empirical Models Example

- As an illustration, consider the data in the table.
- In this table y is the purity of oxygen produced in a chemical distillation process, and x is the percentage of hydrocarbons that are present in the main condenser of the distillation unit.

Hydrocarbon level (X)	Purity (Y)
0.99	90.01
1.02	89.05
1.15	91.43
1.29	93.74
1.46	96.73
1.36	94.45
0.87	87.59
1.23	91.77
1.55	99.42
1.4	93.65

Reference: Applied statistics and probability for engineers, Douglas C. Montgomery, George C. Runger, John Wiley & Sons, 2007



4

We will see one example. There is a table is given there is an X variable called hydrocarbon level Y Variable is called purity as an illustration considered data in the table. In this table Y is the purity of oxygen produced in a chemical distillation process and X percentage of hydrocarbons that are present in the main contents of the distillation unit. Now, we are going to see what is the influence of X on purity of oxygen?

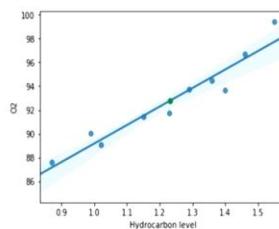
(Refer Slide Time: 02:43)

Using python for plotting the data

```
In [20]: data = pd.read_excel('C:/Users/Somi/Desktop/reg2.xlsx')

In [19]: x= data['Hydrocarbon level']
y = data['O2']
plt.figure()
sns.regplot(x,y,fit_reg= True)
plt.scatter(np.mean(x), np.mean(y), color = "green")
```

Out[19]: <matplotlib.collections.PathCollection at 0x21ada0ab1d0>



5

So, I enter the data in excel so I saved the file name is reg2.xlsx when import data you go to pd.read_excel I have specified the path. So X = data that is a hydrocarbon level that column is my X variable, Y = in data file 'O2' that is my dependent variable and I use plt.figure then sns. regression plot regplot (x, y fit_regression equal to true then can I use this plt.scatter (

`np.mean(X), np.mean(Y), color = 'green')`, would be green so what I am saying and getting a scatter plot between hydrocarbon level and oxygen.

So, what is happening whenever the hydrocarbon level is increasing the oxygen level also increase. There is a positive relationship suppose if you want to make a relation between quantify the magnitude of X and how it is influencing and Y then I should go for regression equation that will do incoming slides.

(Refer Slide Time: 03:58)

Simple Linear Regression Model

- The equation that describes how y is related to x and an error term is called the regression model.
- The simple linear regression model is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where:
 β_0 and β_1 are called parameters of the model,
 ε is a random variable called the error term.



So, the theory behind the simple linear regression model the equation that describes how Y is related to X and an error term is called regression model. The simple linear regression model is $y = (\beta_0) \text{ beta 0} + (\beta_1) \text{ beta 1 } x + (\varepsilon) \text{ error term}$. Here $y =$ equal to $\beta_0 + \beta_1 x + \text{error term}$ where β_0 and β_1 are called the parameter of the model, ε is a random variable called the error term. So what we are going to do we are going to estimate the value of Y with help of independent variable X .

Because X itself will not enough to predict the Y variable there maybe some unknown variable other than X the error due to that unknown variable, otherwise unexplained variance we are going to call it is error term.

(Refer Slide Time: 04:52)

Simple Linear Regression Equation

The simple linear regression equation is:

$$E(y) = \beta_0 + \beta_1 x$$

- Graph of the regression equation is a straight line.
- β_0 is the y intercept of the regression line.
- β_1 is the slope of the regression line.
- $E(y)$ is the expected value of y for a given x value.



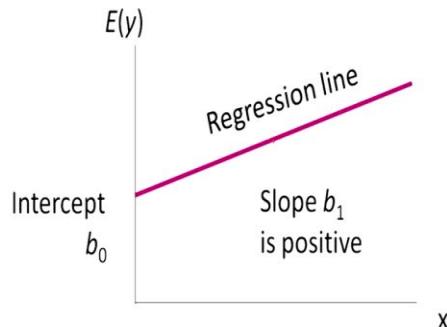
The simple linear regression equation is expectation of Y equal to $\beta_0 + \beta_1 X$ when you comparing the previous slide. That was there is no error term because while calculating the value of β_1 when we have taken care that error is minimized not only that the previously the previous slide we are writing Y now it is expected value of Y . Now what we are predicting is the mean value of Y not the actual value of Y where the graph of the regression equation is a straight line.

Because the power of x is 1, β_0 is the Y intercept of the regression line β_1 is the slope of regression line. So, the expected Y is the expected value of Y for a given X value expected values nothing but mean value.

(Refer Slide Time: 05:41)

Simple Linear Regression Equation

Positive Linear Relationship

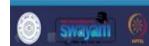
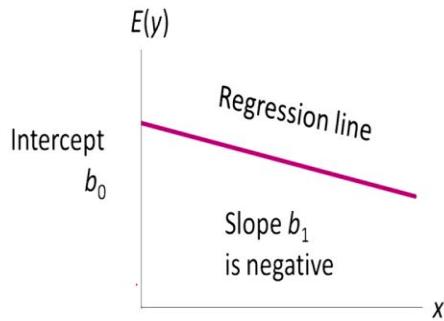


The simple linear regression equations. This is an example of positive linear relationship is in X-axis. There is a when the value of the value of X is increasing the expected value of Y also increasing so the slope beta1 of this b 1 is a positive. The intercept so this distance is your b 0.

(Refer Slide Time: 06:04)

Simple Linear Regression Equation

Negative Linear Relationship

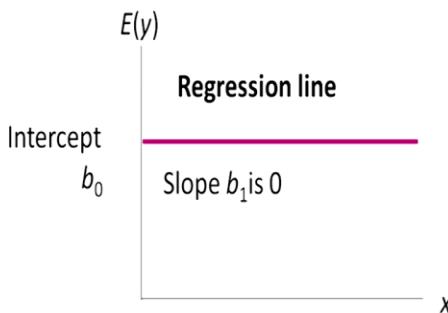


This is an example of negative linear relationship what is happening when X increases the expected value of Y is decreasing here the slope is negative.

(Refer Slide Time: 06:16)

Simple Linear Regression Equation

No Relationship



Here, there is no relationship because it is a line which is parallel to x -axis to the slope is 0. So what is the meaning of this irrespective of any value of X , the expected value of Y is same. Here we can see the value of x and y are independent.

(Refer Slide Time: 06:35)

Estimated Simple Linear Regression Equation

- The estimated simple linear regression equation

$$\hat{y} = b_0 + b_1 x$$

- The graph is called the estimated regression line.
- b_0 is the y intercept of the line.
- b_1 is the slope of the line.
- \hat{y} is the estimated value of y for a given x value.



The estimated simple linear regression equation is $\hat{y} = b_0 + b_1 x$ generally write the capital Y if I write $\beta_0 + \beta_1 X$ if I use capital letters that is for the population. What we write in small letter that is for the sample. So \hat{y} is the estimated regression line b_0 is the y intercept of the line b_1 is the slope of the line \hat{y} is the estimated value of y for given x value.

(Refer Slide Time: 07:09)

Least Squares Method

- Least Squares Criterion

$$\min \sum (y_i - \hat{y}_i)^2$$

where:

y_i = observed value of the dependent variable

for the i th observation

\hat{y}_i = estimated value of the dependent variable

for the i th observation



The principle behind the least square method is the sum of the square of the error has to be minimized. Suppose I have some x value y , I have some number for x I have a number for y suppose, I have drawn this way. This is x axis. This is y axis and plotted line like this. So my objective is I have to draw a line. I have to draw a line. Ideally that line has to pass through all the given points, but that is not possible.

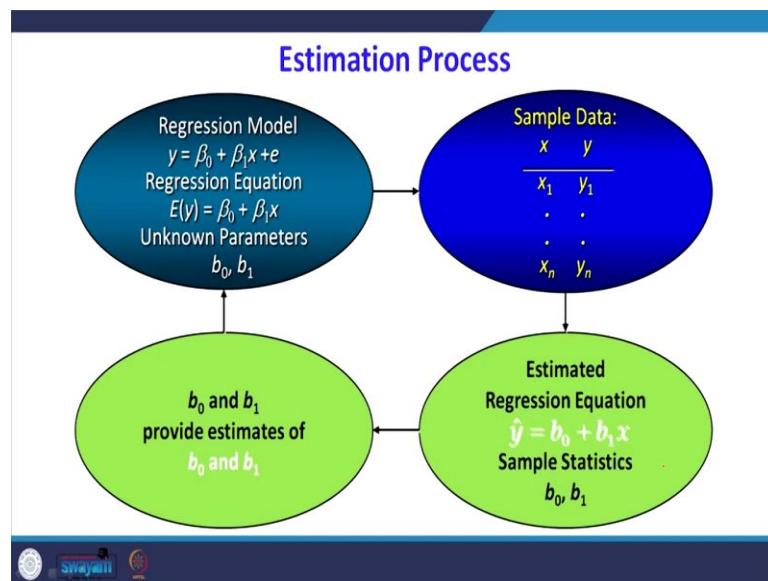
So what I am going to do I am going to draw a line so that the error is minimised not only the for, example this is e_1 this is e_2 this is like that be many e_3 , so this is positive error actual minus predicted value. So this line is $\hat{y} = b_0 + b_1 x$. So now what happening this much distance is it is a vertical lines to this much distance is my error actual minus predicted value. This vertical line distances e_2 so what is happening? This is error.

So what I have to do the error square and sum of the square has to be minimised like this. So, some of the error has to be minimised. I have to draw a line in such a way that sum of the square of the error has to be minimised. I can draw different line suppose. I can draw this, this way also, this way also for each line. I want to find out this sum of the square of error wherever the sum of the square of the error is minimum so that line is the best line that principle called least square method.

Why we are squaring there is logic behind this, if you are squaring the positive and negative error will become nullify we will 0 that is why we are squaring that the same logic for example the formula for variance what we are doing $\Sigma(X - \bar{X})^2 / (n - 1)$, the logical why we ask squaring the 2 purpose otherwise $\Sigma(X - \bar{X})$ equal to 0 here the sum of positive or negative or 0, the square transformation says one more implications that suppose the deviation is less for example it is 0.5 when you square it then it is 0.25.

Suppose deviation is 5 the net value is 55. What is happening? There is lesser deviation had lesser penalty, there is a larger deviation larger penalty. That is beauty of this squared the transformation.

(Refer Slide Time: 10:19)



In the estimation process what is happening initially, we will assume a regression model Y equal to $\beta_0 + \beta_1 X + e$, that regression model will predict with help of regression equation that is expected value of Y equal to $\beta_0 + \beta_1 X$ you see that. The regression equation there is no error term. Here the unknown parameters are the population regression model say Y equal to $\beta_0 + \beta_1 X + e$, the regression equation is expected value of Y equal to $\beta_0 + \beta_1 X$.

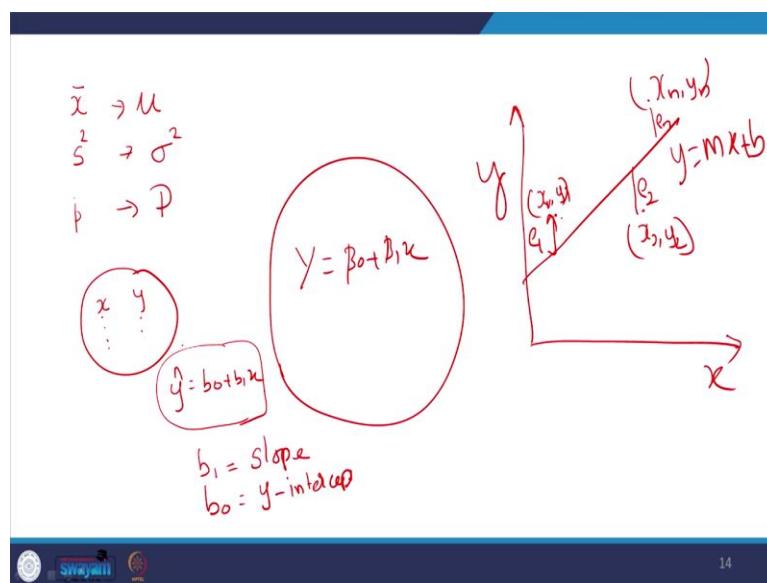
Here the unknown parameters are β_0 and β_1 . We have to estimate the value of β_0 and β_1 more importantly the value of β_1 what you are going to estimate whether the value of β_1 is 0 or other than 0 if I estimate $\beta_1 = 0$ that means there is no relation between X and

Y. So, this is our equation from that what I am going to do I going to collect the sample for my X variable and Y variable X is independent variable. Y is dependent variable. So, with the help of the sample data are going to make a regression equation that is applicable only for the sample. So $\hat{y} = b_0 + b_1 x$, here b_0 and $b_1 x$ is the sample statistics.

So this equation is valid only for the sample now I going to estimate that whether the value of b_1 and b_0 is valid event for the population level also, what is the meaning of that one is sometime this Y you could to $b_0 + b_1 X$ with the help of sample. You can construct a regression equation. What will you predict it when you estimate value of b_1 estimate value of b_0 , then may not be significant other population-level.

So that time what we are going to see the b_1 is equal to 0, if the b_1 is equal to 0 there is no relation between X and Y we will see in the coming slide.

(Refer Slide Time: 12:42)



How this regression is different from our previous concept which have studied. For example what happened with help of \bar{x} we have predicted population parameter mean with the help of sample variance we have predicted the population variance with the help of sample proportion. We have predicted population proportion the regression actually what is happening there is a sample smaller circle sample bigger circle is population.

I have some X and Y value from the sample with help of x and y value I hope predicted regression equation y equal to $b_0 + \beta_1 x$ now I am going to prove that whether this relationship is valid even for the population for that what I am going to do capital $y = \beta_0 + \beta_1 x$ with the help of regression equation. That means a sample model I going to predict whether this model are this relationship is valid for even for the population are not.

Sometime what happened you can construct a regression equation with help of sample data. You can say there is a relation between x and y but when you go to the population level, there were not be relation between x and y. So, what is the difference between this regression modelling? And previous our hypothesis testing, in hypothesis testing we have tested only one parameter at a time what you done we have tested, we are predicted mean are variance are population proportion.

Now I have constructed a model small model with the help of sample data I am testing this model in the population level, I am simultaneously I am checking 2, 3 parameter one is my beta 1 one parameter, beta 0 is another parameter, like that we may have different this is simple regression like that in the multiple regression different independent variable. This is the logic of regression modelling.

What will you do suppose in the regression equation with the help of sample data? What is required is I have to find out what is the value of beta 1, b_1 this is called slope. This is my 'y' intercept what is happening a line is like this suppose there are 2 points. Suppose I am going to call this is a (x_1, y_1) , this is (x_2, y_2) like that there will many point for example this is (x_n, y_n) ok. So this is my x-axis. this is y-axis. So what I am going to do this is my error going to call to e_1 this is e_2 this is e_n so what are you going to do? First go to find out the error for each values then I going to square the error.

Then I am going to sum the error then for what value of this b_1 and b_0 the error will get minimised so that I am doing here the next slide to what is happening. So here I am going to call it is this is y equal to $mx + b$ this is traditional notation because our school in your study this

really you can use any notations. So here what is the error term actual minus predicted so my actual error is for this one my actual point is y_1 my predictable value is y .

(Refer Slide Time: 16:57)

$$\begin{aligned}
 \text{Squared Error (SE)} &= (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \dots + (y_n - (mx_n + b))^2 \\
 &= y_1^2 - 2y_1(mx_1 + b) + (mx_1 + b)^2 \\
 &\quad + y_2^2 - 2y_2(mx_2 + b) + (mx_2 + b)^2 \\
 &\quad + \dots \\
 &\quad + y_n^2 - 2y_n(mx_n + b) + (mx_n + b)^2 \\
 &= y_1^2 - 2x_1y_1m - 2y_1b + m^2x_1^2 + 2mx_1b + b^2 \\
 &\quad + y_2^2 - 2x_2y_2m - 2y_2b + m^2x_2^2 + 2mx_2b + b^2 \\
 &\quad + \dots \\
 &\quad + y_n^2 - 2x_ny_nm - 2y_nb + m^2x_n^2 + 2mx_nb + b^2
 \end{aligned}$$

15

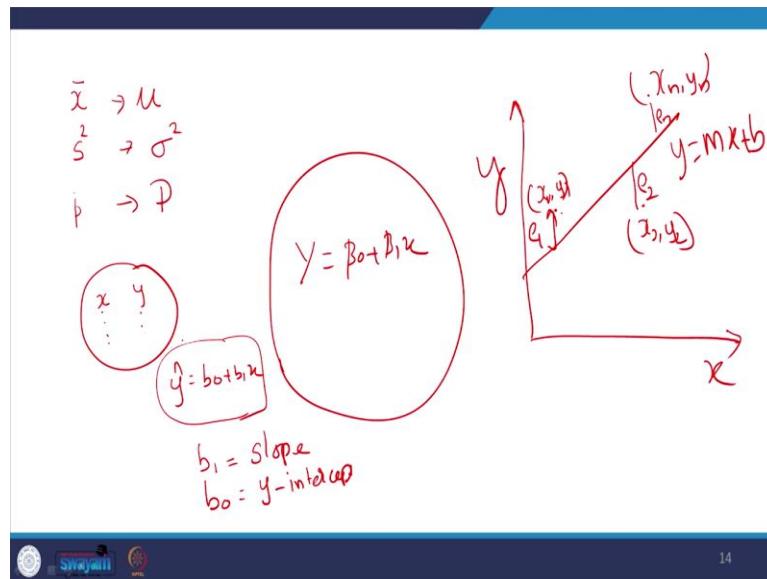
So, $y_1 - y$, y is my $mx_1 + b$ whole square for second term y_2 is actual $mx_2 + b$ because you know this y equal to $mx + b$. so y equal to y_1 when you substitute x_2 equal to x_1 you will get y predicted value actual minus predicted that is a square for finding this actual minus predicted is error and squaring error. so, for the second term $y_2 - mx_2 + b$ like this $y_n - 1 + mx_n + b$ whole square just I am going to simplify this. This is nothing but $a - b$ whole square formula.

y_1 square - $2y_1m$ x_1 b + m x_1 $+ b$ whole square similarly for the pink one y_2 square $a - b$ whole square - $2y_2m$ x_2 $+ b$ + mx_2 $+ b$ whole square for nth term y_m square - $2y_nm$ x_m $+ b$ + mx_b $+ b$ whole square just i am going to simplify this so when you simplify this y_1 square will be there you take this $-2y_1$ inside it will become $-2x_1y_1m$ then $-2y_1b$ + this is $a + b$ whole square, m square xm square + $2x_1b + b$ square this was is the first term.

For the second term y_2 square - $2x_2y_2m$ - $2y_2b$ + m square x_2 2 square + $2m$ $x_2b + b$ square for the nth term y_m square is nothing but sometime $a - b$ sometime $a + b$ whole square formula y_m square - $2x_my_mn$ - $2y_nb$ + m square x_n square + $2x_nb + b$ square. now what i am going to do the next term are going to add all this y_1 square + y_2 square y_n square then here the $-2m$ is the constant here. The next second term so I am going to add this I go to group

it, similarly -2 bs same so I go to add only $y_1 + y_2 + \dots + y_n$ here the m square is the constants $x_1^2 + x_2^2 + \dots + x_n^2$ here 2 mb is a constant so i am going to add $x_1 + x_2 + \dots + x_n$ and here there are b square n time and b square that should have done this one.

(Refer Slide Time: 19:52)



So I have grouped all $y_1^2 + y_2^2 + \dots + y_n^2$ as i told you previously i am going to here the - 2 m is constant so $x_1 y_1 + x_2 y_2 + \dots + x_n y_n$ i may way go back here - 2b for the third term - 2b is the constant. -2 b the remainder is $y_1 + y_2 + \dots + y_n$. the 4th term m square is a constant $m^2 + x_1^2 + x_2^2 + \dots + x_n^2$. then next 2 mb, 2 mb is constant in all the terms so when you bring it common 2 mb the remaining is $x_1 + x_2 + \dots + x_n$ the last term is b square b square b square.

So here but I am going to do you see that. I want to know y^2 mean. so what time to do sigma of $y_1^2 + y_2^2 + \dots + y_n^2$ upto y_n^2 divided by n. so, what are you going to do the submission and go to write in terms of its average. so when i take this one so instead of $y_1^2 + y_2^2$ i can write $n y^2$. similarly this -2 m -2 m so this one i can write $n^2 x^2$ here - 2b what i have done.

i have group the square term. the first term is $y_1^2 + y_2^2 + \dots + y_n^2$ to this i want to write in terms of its average value so i go to write y^2 bar nothing but $y_1^2 + y_2^2 + \dots + y_n^2$ up to y_n^2 divided by n square. Now what is happening Y^2

square y^2 square + yn square can be replaced by multiplied by y square bar. so that is wrote it as y square bar. similarly the second term $x_1 y_1 x_2 y_2 + \dots$ up to $x_m y_n$ can be written as xy bar multiplied by n .

so the remaining term is $2mn$ xy bar the next $- 2b$ that i go to write $n y$ bar. so we will look at the third bar m square, m square will come as it is so x_1 square + x_2 square and so on going to write in terms of n multiplied by x square bar the. Next term is $2mb$ I am going to write $n x$ bar there are $n b$ square writing $n b$ square. So this is the simplified of the error term.

(Refer Slide Time: 23:00)

$$SE = n \overline{y^2} - 2mn \overline{xy} - 2bn \overline{y} + m^2 n \overline{x^2} + 2mbn \overline{x} + nb^2$$

$$\frac{\partial(SE)}{\partial m} = -2n \overline{xy} + 2m \overline{nx^2} + 2bn \overline{x} = 0$$

$$\frac{\partial(SE)}{\partial n} = -2n \overline{xy} + 2m \overline{nx^2} + 2bn \overline{x} = 0$$

$$= -\overline{xy} + m \overline{x^2} + b \overline{x} = 0$$

$$m \overline{x^2} + b \overline{x} = \overline{xy}$$

$$m \frac{\overline{x^2}}{x} + b = \frac{\overline{xy}}{x}$$
one point $(\frac{\overline{x^2}}{x}, \frac{\overline{xy}}{x})$

Actually it is a squared error. So what is happening $n y$ square - n square - $2 m n xy$ bar - $2b ny$ bar + m square $n x$ square bar + $2 m b n x n$ bar b square. Now what is happening here which is variable there are 2 variable one is here comes the slope the slope is the variable because why I am saying slope is variable I can draw different slope different line the slope is variable. So I have to find out for what value of this slow the square of the error will be minimised.

So it is we say it is Maxima minima principal. So what will do for generally what will you do for maxima minima principal dy/dx equal to 0, might have studied in schooling so d square y by d x square is negative less than 0 so that the; it is this way. This way what happening dy/dx equal to 0 is this point if it is negative. That means you are, it is a maximum point if it is a positive. So that is the minimum point.

So, what is happening even to the both for both the conditions. The first one is $\frac{dy}{dx}$ is equal to 0. Here x is variable for example here the m the slope is variable and the b , y intercept is the variable. so, first we will; because there are 2 variable is there you partially differentiate this squared error first with respect to m when you, partially differentiate with respect to m . so there is no n term there will be $-2m\bar{x}\bar{y}$ plus here also there is no m term it is 0 so $2mn\bar{x}^2$ bar here there is x term $2b\bar{n}x$ bar this will become 0 then equate to 0.

So when you simplify this $2n$ is the constant $2n$ here $2n$, $2n$ remove this the remaining $-\bar{xy}$ bar square + $m\bar{x}^2$ bar+ $b\bar{x}$. so i am going to write in this one, y equal to $mx+b$ format. so what will happen $m\bar{x}^2$ + $b\bar{x}$ take right and side \bar{xy} bar, so i am going to divide by \bar{x} bar so \bar{x}^2 by \bar{x}^2 + b . this is \bar{xy} bar – \bar{xy} bar divided by \bar{x} bar now what happening this is $mx+b$ equal to y format which one this equation.

So, m is m so the x coordinate is \bar{x}^2 bar divided by \bar{x} bar so the y coordinate is \bar{xy} bar divided by \bar{x} bar. so, what this implies is if you want to draw a best line that line has to pass through this point this is x coordinate this one is this is first one x coordinate and the second one y coordinate. so if you want to draw a line which should minimise the sum of the squared error that has to pass through this point.

(Refer Slide Time: 26:27)

$$\begin{aligned}
 SE &= n \bar{y}^2 - 2mn\bar{xy} - 2bn\bar{y} + m^2n\bar{x}^2 + 2mbn\bar{x} + nb^2 \\
 \frac{\partial(SE)}{\partial b} &= -2n\bar{y} + 2mn\bar{x} + 2nb = 0 \\
 &= -\bar{y} + m\bar{x} + b = 0 \\
 \bar{y} &= m\bar{x} + b \\
 \text{another point } (\bar{x}, \bar{y}) &
 \end{aligned}$$

Then we will find out the other because to know the slope we need 2 point we got already one. So now, we differentiate partially differentiate with respect to b so here there is no b term 0, here also, there is no b term 0 here there is a b term -2 my bar there is no b term here $2mn \bar{x}$, so here $2nb$ so equate to 0 here also this $2n$ is a constant $2n$ so divide both side remaining $-\bar{y}$ + $m\bar{x} + b$ equal to 0 so when you simplify y equal to $mx + b$ format.

so now this is y equal to $mx + b$ format this line passing through the \bar{y} so this line is passing through \bar{x} so the another point is \bar{x}, \bar{y} so you see this is very important result. if you want to draw a best line that line has to pass through the average value of its x and average value of y , one of the point should that lines to pass through that then only that line maybe the best line. so we got the 2 point \bar{x}, \bar{y} another point is $\bar{x^2}/\bar{x}, \bar{xy}/\bar{x}$ divided by \bar{x} .

so, when we know this one point say another point is \bar{x}, \bar{y} so if you we want to know the slope of this equation then what is the slope formula $(\bar{y} - \bar{y})/(x - \bar{x})$ when you use that formula you will get formula for slope. dear students we got the 2 points after using the least square principle the one point is $\bar{x^2}/\bar{x}, \bar{xy}/\bar{x}$ divided by \bar{x} that is one point.

Another point is \bar{x}, \bar{y} when there are 2 point is there we can find out the slope. what is the slope formula of this is first point, what is the slope formula suppose a traditional we might have studied this in school. Suppose there is 2 points point 1 is x_1, y_1 so point 2 is x_2, y_2 ok. so, that x point is $x^2/\bar{x}, y_1$ point is divided by xy/\bar{x} divided by \bar{x} . so, x_2 point is $x/\bar{x}, y/\bar{x}$ we know the slope formula $(y_2 - y_1)/(\bar{x} - x_1)$. here the y_2 is $y/\bar{x} - y_1$ xy/\bar{x} divided by \bar{x} divided by \bar{x} is x/\bar{x} minus actually this is i wrote x_1, y_1 for only our convenience.

so, this x_1 is different x_2 is $x/\bar{x}, x_1$ is $x^2/\bar{x}, x^4/\bar{x}$. so this is when you multiply both side numerator and denominator by x/\bar{x} it will become $x/\bar{x}, y/\bar{x} - xy/\bar{x}$ divided by x/\bar{x} whole square, x^2/\bar{x} because it is $x/\bar{x} x/\bar{x}$ get cancelled when you bring minus this one when you multiply both side by minus xy/\bar{x} square - $x/\bar{x}, y/\bar{x}$ divided by x^2/\bar{x} -

$x \bar{x}$ whole square. this is the formula for slope actually this slope is nothing but when you look at the numerator, there is nothing but the covariance of (x, y) the denominator is variance of x i will explain how this numerator is covariance.

so, we know that in our probability class you study the covariance of x, y is expected value of $x - x \bar{x}$ by $y - y \bar{y}$ and you simplify this you bring this side $xy - xy - xy \bar{x} - x \bar{x} y + x \bar{x} y \bar{y}$ we you bring e inside? it will become $e x y$ because $y \bar{y}$ is number when you bring your e of x again it is $x \bar{x} - x \bar{x}$ is a number when you bring e inside it will become $y \bar{y}$ because $y \bar{y}$ is a number so that will be as it is. so, when you bring e here it becomes $xy \bar{x}$ so $- x \bar{x} y \bar{y} - x \bar{x} y \bar{y} + x \bar{x} y \bar{y}$ you can cancel it plus and minus.

the reminder is $xy \bar{x} - x \bar{x} y \bar{y}$ that is nothing but the numerator and going back you see that $xy \bar{x} - x \bar{x} y \bar{y}$ is numerator. so, that in the slope formula numerator is nothing but covariance of x, y so the variance of x is this also we studied from school $x - x \bar{x}$ whole square you expand $x^2 - 2x \bar{x} + x \bar{x}^2$ when you bring x inside this will become e of $x^2 - 2x \bar{x}$ is number the expected value of a number is number itself and , d of x becomes $x \bar{x}$ this is number itself will keep as it is.

when we keep e inside become $x^2 - x \bar{x}^2$ so there are $- 2x \bar{x}$ whole square + $x \bar{x}$ whole square. so when you subtracted it the remaining is 1. so, $x^2 - x \bar{x}^2$ even when i go back when we look at this, this formula is sampling with this denominator. so, the slope formula the numerator is nothing but write numerator is nothing but the covariance denominator nothing, but the variance of x .

Actually this variance covariance and correlation coefficient regression slope all are having some relationship. See that the variance formula we know sigma of $x - x \bar{x}$ the whole square divided by $n - 1$ that is only one variable. in the covariance there are 2 variable x, y , so sigma of $x - x \bar{x}$ even this can be written as sigma of $x - x \bar{x}$ into $x - x \bar{x}$ there 2 variables is there instead of another $x - x \bar{x}$ you can write sigma of $x - x \bar{x}$ into $y - y \bar{y}$ divided by $m - 1$.

so, the correlation coefficient is nothing but when you divide this covariance divided by its own standard deviation, it will get correlation coefficient. so, but the slope is the ratio of covariance divided by variance of x you see the variances. the covariance is this 1 sigma of $x - \bar{x}$ divided by $y - \bar{y}$ by $n - 1$ assume that the equal same degrees of freedom. the denominator is sigma of $x - \bar{x}$ bar of whole square $n - 1$. so, this $n - 1$ get cancelled the formula for slope is sigma of $x - \bar{x}$ bar into $y - \bar{y}$ bar divided by sigma of $x - \bar{x}$ bar whole squared.

(refer slide time: 33:14)

Least Squares Method

- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

that is why we got this formula b_1 so we need not it is easy way to remember this formula for the slope is nothing but covariance by variance. in this class we started about the regression analysis, I have explained the importance of regression, how the regression is different from the our traditional hypothesis testing. in the regression equation there is a y intercept is there and slope there by using least square method. i have derived the formula for finding the slope and the y intercept.

The formula for slope is nothing but the covariance by variance then I have interlinked how variance covariance and correlation coefficient and regression coefficients. All these are interrelated. The advantage here is you need not remember the formula, formula is; if you know the variance formula and covariance formula easily, you can find out the slope of regression equation. We will continue the next class by taking an example; I will explain how to use this regression equation for prediction purpose. Thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 29
Linear Regression - II

Dear students in the previous class I have derived what is the formula for Slope of a linear regression equation and y intercept and also I have explained the concept of least square method. In the formula of slope then I explained slope is nothing but covariance of the two variable if it is the 1 independent variable and one dependent variable simple linear regression. Covariance divided by variance, variance of independent variable.

Then we also got another important result in the previous class that if you want to draw a best line that line has to pass through its average of x, y value that that lines to pass through average value of x that x bar and average value of y bar class what you are going to do.

(Refer Slide Time: 01:22)

Least Squares Method

- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

 IIT Roorkee

I have taken the problem with help of a small problem I going to find out the slope and y intercept then I will explain the practical meaning of this y intercept and slope this formula b1 nothing, but the slope of the line is writing $y = b_0 + b_1 x$ to the b1 is nothing but the slope is covariance divided by variance, we know that the formula for covariance is $\Sigma (x - x \bar{x}) . (y - y \bar{y})$ divided n - 1 divided by the variance.

The variance is $\Sigma (x - \bar{x})^2 / (n - 1)$ because the numerator and the denominator $n - 1$ is same so when you cancel that the remaining formulas, $\Sigma (x - \bar{x}) . (y - \bar{y})$ divided by $\Sigma (x - \bar{x})^2$.

(Refer Slide Time: 02:22)

Sum of squares and sum of cross-products

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y})$$

 Swayam 

3

If you are using calculator for examination purpose this is the very very very useful notations convention. So by using if you know S_{xx} the meaning of this is S_{xx} is $\Sigma (x - \bar{x})^2$ the meaning of S_{yy} is, it is a convention is $\Sigma (y - \bar{y})^2$ if I write S_{xy} that is nothing but S_{xy} equal to $\Sigma (x - \bar{x}) . (y - \bar{y})$.

(Refer Slide Time: 03:01)

Sum of squares and sum of cross-products

$$Slope(m) = \frac{S_{xy}}{S_{xx}}$$

$$SSE = \text{error sum of squares} = S_{yy} - \frac{S_{xy}}{S_{xx}}$$

 Swayam 

4

The formula for slope is nothing but slope m equal to S xy divided by S xx. If I want to know error sum of square, I will explain what is the meaning of error sum of square. Now you take this formula what is error sum of square equal to S yy – (S xy divided by S xx) suppose why this formula so useful suppose if you know this 3 term S xx, S yy, and S xy you can find out the slope. And you can find out the error sum of square and this is convenient also later to find out the coefficient of determination. I will explain the next lecture.

(Refer Slide Time: 03:52)

Simple Linear Regression

- Suppose that we have n pairs of observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.
- Previous Figure shows a typical scatter plot of observed data and a candidate for the estimated regression line.
- The estimates of β_0 and β_1 should result in a line that is (in some sense) a "best fit" to the data.
- The German scientist Karl Gauss (1777–1855) proposed estimating the parameters β_0 and β_1 in Equation

$$\hat{y} = b_0 + b_1 x + \varepsilon$$

to minimize the sum of the squares of the vertical deviations

So what is simple linear regression suppose that we have n pairs of observation se(x1, y1),(x2, y2) like this (xn, yn).

(Refer Slide Time: 04:07)

Least Squares Method

y-Intercept for the Estimated Regression Equation

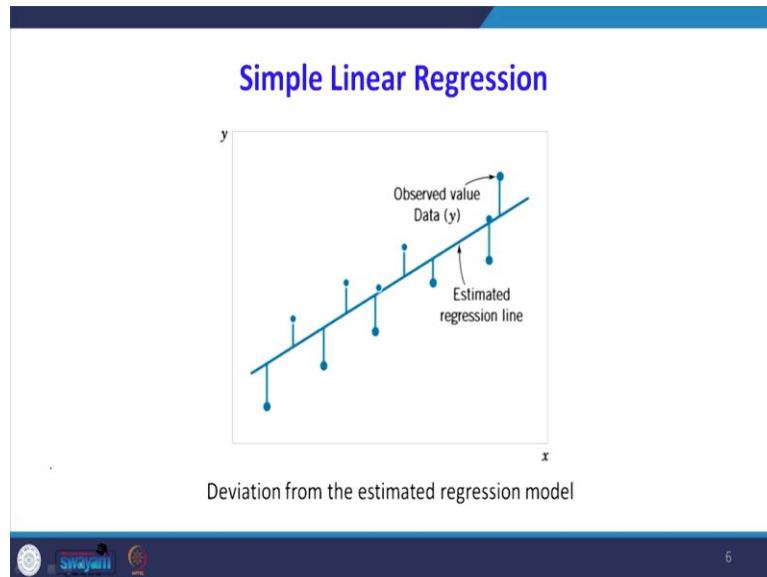
$$b_0 = \bar{y} - b_1 \bar{x}$$

where:

- x_i = value of independent variable for i th observation
- y_i = value of dependent variable for i th observation
- \bar{x} = mean value for independent variable
- \bar{y} = mean value for dependent variable
- n = total number of observations

Previously we have seen the formula for slope. Now the formula for y- intercept to be zero equal to $y_{\bar{}} - b_1 \bar{x}$, explain how this formulas come here x_i is value of independent variable for the i th observation, y_i is the value of dependent variable for right observation \bar{x} is the mean of mean value for independent variable, $y_{\bar{}}$ is mean value of value for dependent variable n actually we not using here n is the total number of observations to find out the mean value.

(Refer Slide Time: 04:46)



Simple Linear Regression is you this is your x axis. This is your y -axis. So whatever value which is said that is observed value actual value this line shows the critical value. So this line generally written as $b_0 + b_1 x$ the final objective in a regression equation is to find out what is the slope of this line and y intercept of a line because if you know slope and y intercept, so this was your y intercept, if you know, y intercept and slope then you can construct the regression equation.

The previous class already explain this is your error one. This is error 2 this is error 3 so the concept of least square method is the sum of the square of the error has to be minimised. So that idea is taken care to find out the value of slope and y intercept.

(Refer Slide Time: 05:57)

Simple Linear Regression

Example: Auto Sales

An Auto company periodically has a special week-long sale.

As part of the advertising campaign runs one or more television commercials during the weekend preceding the sale.

Data from a sample of 5 previous sales are shown on the next slide.



We will take one example simple linear regression why it is called simple linear regression only one independent variable is there. If there are more than one independent variable that you will call it as multiple linear regression small problem and explain how to construct a regression equation and how to use this formula of slope and y intercept. An auto Company periodic special week-long sale as a part of the advertising campaign Company runs one or more television commercials during the weekend preceding the sale.

Data from a sample of 5 previous sales are shown in the next slide actually the company before introducing a new product. They go for television advertisement. This problem says, is there any effect of television advertisement on the sales of the car?

(Refer Slide Time: 06:56)

Simple Linear Regression

Example: Auto Sales

<u>Number of TV Ads</u>	<u>Number of Cars Sold</u>
1	14
3	24
2	18
1	17
3	27



The data says number of TV ads 1 the number of cars sold 14 when the number of TV ads is 3 number of cars sold is 24, number of TV ads is 2 number of cars sold is 18 number of TV ad is 1 number of cars sold is 17 number of TV ads is 3 number of cars sold 27. In this the dependent variable is number of cars sold this generally we will call it as y is dependent variable. The independent variable is x that is nothing but number of TV ads.

So we have to know effect of this number of ads on the number of cars sold. Generally what is perception when you the frequency of ads is more than be more sales.

(Refer Slide Time: 07:51)

Estimated Regression Equation

Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{20}{4} = 5$$

y-Intercept for the Estimated Regression Equation

$$b_0 = \bar{y} - b_1 \bar{x} = 20 - 5(2) = 10$$

Estimated Regression Equation

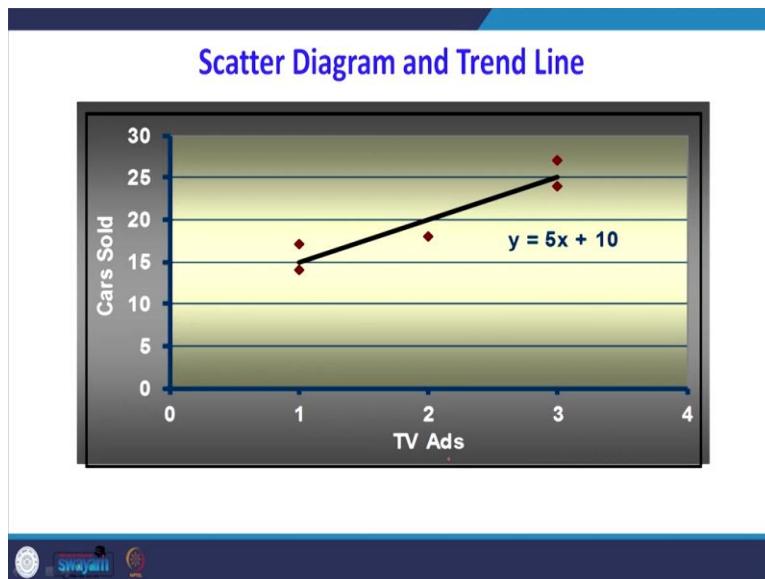
$$\hat{y} = 10 + 5x$$



The first task is the Slope of the estimated regression equation. So what we have to do first we have to find out \bar{x} and \bar{y} then for each value of x and find out $x - \bar{x}$ and for each value of y you find out $y - \bar{y}$ then you have to multiply that and you have to sum that multiplication that lead to 20. Then sigma of $x - \bar{x}$ you have to square that then you to submit that is 4. So the slope is 5 the y intercept for estimated regression equation is b_0 is $\bar{y} - b_1 \bar{x}$ already we know b_1 .

So you take b_1 value here \bar{y} you know it is 20 the \bar{x} bar is 2 this is 10 so the estimated regression equation is $10 + 5x$, you are to be very careful. This is estimated regression equation. It is not why the value of y is the estimated value it is not the actual value it is nothing but the mean value. So, y equal to $10 + 5x$ so how to interpret this, when the value of x is increasing 1 unit y will be increased by 5 units so keeping other things constant when x is increased by 1 unit the sales will increase by 5 times not the 15 times it is not right 5 into 1 equal to 15. We have seen the rate of increment of x and rate of increment of y .

(Refer Slide Time: 09:35)



So, that this show that $y = 5x + 10$. So here the 10 is y intercept when you extend this. This is a y intercept. Ok 5 is the slope. So suppose if the TV number of TV Advertisement is a 5 now we have taken up to 4 suppose this is 5 you can put to here $5, 25 + 10, 35$. This way; this is a trend line is a regression line.

(Refer Slide Time: 10:08)

Jupyter Code

```
In [2]: import numpy as np  
import matplotlib.pyplot as plt  
  
In [3]: import seaborn as sns  
  
In [4]: import pandas as pd  
import matplotlib as mpl  
import statsmodels.formula.api as sm  
from sklearn.linear_model import LinearRegression  
from scipy import stats  
  
In [5]: tbl = pd.read_excel('C:/Users/Somi/Documents/regr.xlsx')
```

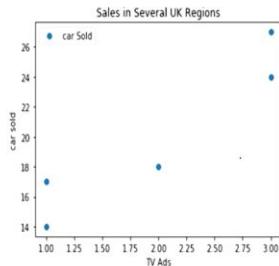
11

This one will do with the help of python import numpy as np import matplotlib.pyplot as plt import seaborn as sns import Pandas as pd import matplotlib as mpl import statsmodels.formula.api as sm, from sklearn and see that this one is sklearn that is the library for running linear regression sklearn.linear_model import linear regression, from scipy import stats first I have entered this value in excel and going to save that filename object called tb1, tb1 equal to pd.read_excel. This was the path where I have stored my excel file.

(Refer Slide Time: 11:16)

Jupyter Code

```
In [6]: tbl.plot('TV Ads', 'car Sold', style='o')  
plt.ylabel('car sold')  
plt.title('Sales in Several UK Regions')  
plt.show()
```



12

First task is your to plot the scatter plot scatter plot is we have to see only we can go for correlation here. Also, we can discuss scatter plot will say rough idea about what will happen when the value of x increases and how it is affected y what is happening? You see that you want

some point when the number of TV ads increasing the car sales also increasing so `tbl.plot('TV ads','cars sold', style = 'o')`, `plt.ylabel(" cars sold")`. `Plt.title(' sales in UK regions')`, `plt.show()`. The that will show the this graph.

(Refer Slide Time: 12:01)

```
In [5]: t = tbl['TV Ads']
c = tbl['car Sold']

In [8]: import statsmodels.api as s
t = s.add_constant(t)
model1 = sm.OLS(c,t)
result1 = model1.fit()
print(result1.summary())
```

OLS Regression Results					
Deg. Variable:	car Sold	R-squared:	0.877		
Model:	OLS	Adj. R-squared:	0.836		
Method:	Least Squares	F-statistic:	21.43		
Date:	Fri, 30 Aug 2019	Prob (F-statistic):	0.0190		
Time:	08:31:20	Log-Likelihood:	-9.6687		
No. Observations:	5	AIC:	23.34		
DF Residuals:	3	BIC:	22.56		
DF Model:	1				
Covariance Type:	nonrobust				
	coef	std err	t	P> t	[0.025]
Const	10.0000	2.366	4.236	0.024	2.469
TV Ads	5.0000	1.800	4.629	0.019	1.563

Omnibus: nan Durbin-Watson: 1.214
Prob(Omnibus): nan Jarque-Bera (JB): 0.674
Skew: 0.256 Prob(JB): 0.714
Kurtosis: 1.276 Cond. No. 6.33

13

Next one I am going to save that TV ads in variable quantity equal to `t = tv ads` in and `c = car sold` here. The car sold is dependent variable TV ads is independent variable import `statsmodels.api as s`, so `t = s.add_constant(t)` because we need to have the `t` so `model1` I am saying `model1 = sm.OLS(c,t)`, ols means ordinary least square method `c, t`. What is `c` here? `c` is your dependent variable `t` is your independent variable `result1` equal to `model1.fit()` so `print(result1.summary())`.

This is was the output of your linear regression equations. So, look at this most importantly is it that the coefficient so how to write this one `y` equal to so the constant is $10 + 5 \text{ TV ads}$. There are many terms are here model is ols methos is least square when it was conducted number of observations 5 number of residuals residual is nothing but you error here your r square is 0.877, I will explain the meaning of 0.877 in the next class.

Then the adjusted r square this value of adjusted r squared interpreted for multiple regression equation. I will explain what is this F statistics in next class then there are many fitness index is there. So these standard error this is the t value I will explain t value and one more thing there in

look at the probability value suppose Alpha equal to 5 percentage the probabilities less the 0.05 then you can say it is significant I will explain this also in the next class. So this is the output of our regression equation. We will take another problem on regression analysis.

(Refer Slide Time: 14:33)

The slide has a blue header bar. Below it, the title 'Example Problem- II' is centered in a blue box. The main content area contains a bulleted list of tasks and a reference note at the bottom right. At the very bottom, there is a dark blue footer bar with three small icons on the left and the number '48' on the right.

Example Problem- II

- The data in the file hardness.xls provide measurements on the hardness and tensile strength for 35 specimens of die-cast aluminum.
- It is believed that hardness (measured in Rockwell E units) can be used to predict tensile strength (measured in thousands of pounds per square inch).
 - a. Construct a scatter plot.
 - b. Assuming a linear relationship, use the least-squares method to find the regression coefficients b_0 and b_1 .
 - c. Interpret the meaning of the slope, b_1 , in this problem.
 - d. Predict the mean tensile strength for die-cast aluminum that has a hardness of 30 Rockwell E units.

Reference: Applied statistics and probability for engineers, Douglas C. Montgomery, George C. Runger, John Wiley & Sons, 2007

The problem is the data in the file. I have a file called hardness.xlsx, provides measurement on the hardness and tensile strength for 35 specimen of die cast aluminium. It is believed that hardness that is measured in Rockwell E unit can be used to predict the tensile strength measured in 1000 of Pounds per square inch. So, what are the things used to do is construct a scatter plot assuming a linear relationship. Use the least square method to find the regression coefficient for b_0 and b_1 interpret the meaning of slope b_1 in this problem.

Predict the main tensile strength for the die cast aluminium that has hardness of 30 Rockwell E unit. Today is a tensile strength is given hardness is given for this data set. We are going to construct a regression equation. So will switch to Python I will tell you how to do that.

(Refer Slide Time: 15:27)

```

In [1]: import pandas as pd
import numpy as np
from sklearn import linear_model
import statsmodels.api as sm
from sklearn.metrics import mean_squared_error

In [2]: data = pd.read_excel("HARDNESS.xls")

In [3]: from sklearn.model_selection import train_test_split

In [4]: x = data['Hardness'].values.reshape(1,1)
y = data['Tensile strength'].values.reshape(-1,1)
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.2,random_state = 88)

In [5]: x_train.shape, x_test.shape, y_train.shape, y_test.shape

In [6]: len(x_train)

```

For this import pandas as pd, import numpy as np, from sklearn import linear_model, import statsmodels.api as sm and from sklearn.matrix import mean_squared_error. first I will load the data object called data. So run It import pandas as pd, import numpy as np, from sklearn import linear_model, import statsmodels.api as sm and from from sklearn.matrix import mean_squared_error the file I have stored in a object called data.

So the data source this is the tensile strength and hardness. What are going to do? There are 35, data set that we are going to split into two categories only for training and other only for testing. For that purpose from sklearn.model_selection import train_test_split, x equal to data so x value is going to be hardness dot values dot where reset command is used to convert one dimensional array into 2 dimensional array, then y equal to data that is a tensile strength.

So, x is independent variable, y is dependent variable. The next one is x underscore train, x underscore test, y underscore train, y underscore test equal to train underscore test underscore split x, y test underscore size equal to 20%, so this 20% What is the meaning is there 20% the data will be kept for testing our model remaining 80% of the data will be used for building our regression model. So random underscore stat equal to you can give any number so that when you repeat this program again, you will get the same answer.

Because the 20% of the data is randomly chosen out of 35 data set. So if you use this shape command now you can see I will run this one, so you are getting 28, 1, so 28 data sets for training 7 data status for testing ok. So, we can see the length also also can see that the 28 for training 7 is for testing this is the train data sets. Now will go for constructing the regression model from a scale and linear underscore model input linear regression, so regression equal to Linear regression when you run they will run this model.

So now we will see what is y interested? Now y intercept is 7.045 so the regression coefficient is 1.9974. Now will predict for the test data set will predict what is the y value? So this is your predicted y value when giving x data set as an input. Now we will find out what is the mean square error the mean squared error, the mean squared error is 35, if the error is smaller that model is good model. The next one is the fitness of this regression model is nothing 0.53 by taking x underscore data set an independent variable and y underscore data set is dependent variable. The next is when you set for training data set it is 0.45 explain the meaning of this score data.

(Refer Slide Time: 19:21)

```

Score = R^2 =  $\frac{SSR}{SST}$ 
SSE / n-2

In [30]: mean_squared_error(y_test, y_predict)
Out[30]: 35.71053398209997

In [31]: reg.score(x_test, y_test)
Out[31]: 0.5362243730094254

In [32]: reg.score(x_train, y_train)
Out[32]: 0.4500146647765303

```

This the meaning of score is meaning of score is nothing but your r square. This r square is nothing but coefficient of determination. So, the coefficient of determination is your SSR divided by SST, SSR is regression sum of square, SST is total sum of square in your problem the r square is see that the reg.score (x_test , y_test) is the r square is 53%. What is meaning of this 53

is the 53% of the variable variability of y is explained with the help of this is dependent variable for the training data set.

It is for the training dataset it is 0.45. that is the 45% of the variability of y is explain with help of independent variable x this mean square mean squared error is nothing but SSE divided by $n-2$ that is mean square error. If it is the lesser value the model is good fit. Otherwise, it is not good. Now will see the another concept in the regression model that is called machine learning in machine learning this is one category of supervised learning. The machine learning techniques are classified into two categories under supervised learning method and unsupervised learning method.

So the regression is example for supervised learning because in advance we are labeling what is independent variable and what is dependent variable if it is unsupervised learning. So, we cannot label in advance that what is going to be dependent and what is independent variable. So, in the supervised learning is nothing but the regression analysis. That in the context of machine learning will call it is supervised learning in statistics will call it is simple regression.

Dear students in the previous class derive the formula for y intercept and slope. Then I have taken one sample problem with the help of sample problem explain how to use that formulas and also called with the help of python. I have taken another problem also in that problem. The data set is divided into two parts, 1 part is for Training the for building the data set the using training data set the other part of the data set that is for test data set.

So the test data set was used for validating the model which we have constructed. Thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 30
Linear Regression – III

Dear students in the previous class I would take in a sample example I have explained to construct how to construct a regression equation.

(Refer Slide Time: 00:37)

Learning Objectives

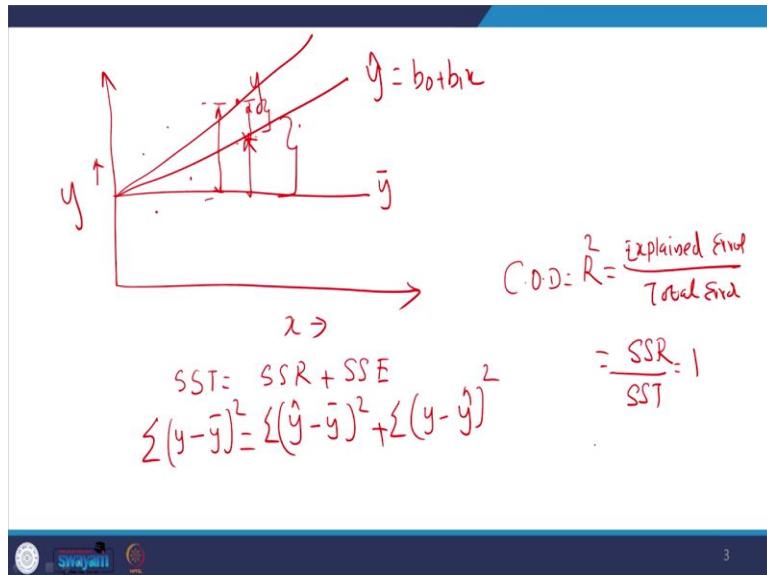
- Understanding Coefficient of Determination
- Test statistical hypotheses and construct confidence intervals on regression model parameters



2

In this class I will explain what is the meaning of coefficient of determination and test statistical hypothesis and construct confidence interval and regression model parameters, there are one parameter is the ‘b’ the coefficient of x is one parameter.

(Refer Slide Time: 01:01)



Now I will explain what is the goodness of it that is coefficient of determination r square assume that why value assume that there is no independent variable the easiest way for prediction is mean of the y what is the meaning of there is no independent variable. Suppose I have demand for say first eleven month I want to predict in 12th month what is the demand of a particular product, the easiest way is you find the mean of the previous data.

So that will be used as the mean for the next data so without considering any independent variable suppose if there is no independent variable so the actual point is say this is the y is the actual value. So, the one way to predict without any independent variable is say y bar but I know there is a one independent variable what is this much this is actual this is predicted, so this much is my error what is this error this point to this point this is error.

What is the error total error we can say total error actual - predictor. Suppose if I know one independent variable that I have as I am assuming that is affecting my dependent variable then that regression equation is like this, so this I am writing $\hat{y} = b_0 + b_1 x$, now what has happened now this much distance because this point so this much distance is see total error is this much, so this much error is you could draw it this way.

So here what is yes this much portion this much portion I am able to explain with the help of regression independent variable. So, total error is this point to this point so this much error with

the help of independent variable x I am able to explain. So, the remaining error this one's this much distance is unexplained the error. So, what I am saying this is only one point there is no linear relationship like that there are different y values may be why here y may be here y may be here maybe here maybe here.

So if I find the total sum of square there is a total error then nothing but y SST . So, SST equal to SST + SSE what is the SST total sum of square what is the SSR a regression sum of square what is SSE error sum of square. So, now what is the logic behind this is the total error is this point to this point total error I am splitting that error due to how much error we are able to explain with the help of this independent variable x that is SSR, so the remaining portions that is the which is in the bracket that is unexplained error.

You when you look at this there will be a connection with on over in ANOVA what do I written the same thing you know what you written SST equal to SS treatment + SSE , your treatment is nothing but your independent variable . Now we will find out what is the formula this so what is SST so this point is y the total error $\Sigma (y - y \bar{ })^2$, that is SST equal to so this much portion what is the error is $\Sigma(y \hat{ } - y \bar{ })^2 + SSE$ unexplained error.

What is unexplained error $\Sigma (y - y \hat{ })^2$,so what has happened the total error is the regression sum of square + error sum of square. Here I want to predict the coefficient of determination, the coefficient of determination referred as r square is nothing but explain the error divided by total error what is explained error explained is nothing but SSR that means that much error we are able to explain because that much error is due to this independent variable x what is a total error this is SST total sum of square.

There is a two possibility of this r square is it cannot be more than 1, if it is 1 what is the meaning the total SST the numerator also SST denominator also SST so what we are saying this point this line pass through that point. If it is less than 1 so what is happening SSE error is smaller SST is bigger, if equal to 1 both are same. So, the upper limit of the r square is 1 the lower limit is 0 to 1 so 0 to 1 is the interval for r square.

(Refer Slide Time: 07:32)

Coefficient of Determination

- Relationship Among SST, SSR, SSE

$$SST = SSR + SSE$$

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$$

$$SS_{yy} = \left(\frac{SS_{\hat{y}\hat{y}}}{SS_{xx}} \right) + \left(SS_{\hat{y}\hat{y}} - \frac{SS_{\hat{y}\hat{y}}}{SS_{xx}} \right)$$

where:

SST = total sum of squares

SSR = sum of squares due to regression

SSE = sum of squares due to error



We will see this one yes, see the relationship among SST, SSR and SSE. So, SST you see $\Sigma (y - \bar{y})^2$ equal to SSR $\Sigma (\hat{y} - \bar{y})^2$ for SSE $\Sigma (y - \hat{y})^2$ remember the previous also I was showing this SS yy that that SS yy is nothing but $\Sigma (y - \bar{y})^2$ so, this is a very handy formula if you are using calculator and a very short cut very quickly you can get the answer for what is SST, SSR and SSE from that you can easily find out r square nothing but SSR divided by SST.

(Refer Slide Time: 07:58)

Coefficient of Determination

- The coefficient of determination is:

$$r^2 = SSR/SST$$

where:

SSR = sum of squares due to regression

SST = total sum of squares



R square is SST / SST, SSR is sum of square due to regression SST is total sum of square.

(Refer Slide Time: 08:07)

Jupyter code

In [5]: `t=tbl[['TV Ads']]
c=tbl[['car Sold']]`

In [8]: `import statsmodels.api as s
t = s.add_constant(t)
model1 = sm.OLS(c,t)
result1 = model1.fit()
print(result1.summary())`

OLS Regression Results

Dep. Variable:	car Sold	R-squared:	0.877			
Model:	OLS	Adj. R-squared:	0.836			
Method:	Least Squares	F-statistic:	22.49			
Date:	Fri, 30 Aug 2019	Prob (F-statistic):	0.0198			
Time:	08:31:28	Log-Likelihood:	-9.6687			
No. Observations:	5	AIC:	23.34			
Df Residuals:	3	BIC:	22.56			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	10.0000	2.366	4.226	0.024	2.469	17.531
TV Ads	5.0000	1.800	4.629	0.019	1.563	8.437
Omnibus:	nan	Durbin-Watson:	1.24			
Prob(Omnibus):	nan	Skew:	0.674			
Kurtosis:	1.276	Prob(OB):	0.714			
		Cond. No.	6.33			

You see that the previous also I was saying what is the meaning of this r square. So, we are getting in our problem we are getting 0.87 I will show you how it has come yes the coefficient of determination formula is r square equal to SSR divided by SST, so SSR in our problem is 100 SST is 114 so I can show you this how this SSR 100 you see that SSR.

(Refer Slide Time: 08:42)

Coefficient of Determination

$$r^2 = \text{SSR/SST} = 100/114 = .8772$$

The regression relationship is very strong; 88% of the variability in the number of cars sold can be explained by the linear relationship between the number of TV ads and the number of cars sold.

So how what is the formula for SSR we have seen previously how we are getting SSR you see that SSR, is nothing but see $\Sigma (y\hat{} - y\bar{})^2$ square first you have to find out the regression equation in that when you substitute first value of x you will get $y_1\hat{}$, $y_1\hat{}$ is when substitute the value of x into that then $y - y\bar{}$ whole square then when you substitute x equal to 2 they will get $y_2\hat{}$ so then $y\bar{}$ whole square when you sum that one that is nothing but your SSR.

SST is $y - \bar{y}$ it is a numerator of that variance of $y - \bar{y}$ whole square so from that you can find out SST. So, in this our problem it is 100 order by 114 now what is the meaning of this r square so the meaning of r square is as we know that it is 0 to 1 the regression relationship is very strong what is the meaning is 88% of the variability in the number of cars sold can be explained by the linear relation between the number of TV ads and the number of cars sold.

So what is meaning that 87 it is 87.7, 88% of the variability of y can be explained by the help of this independent variable there is a remaining 13% that we are not able to explain that may be due to two reasons one is we might all miss you that some other independent variable there may be some other variable that affects the car sales. Another reason is that we have fixed a linear regression but the actual data may follow non linear regressions so that is why we are not getting exactly 1.

In Python output you see that when you see the r square is the 0.77 this is r square is 0.77 that is the meaning of that is 87.7% of the variability of car sold can be explained with the help of number of TV ads that is our independent variable.

(Refer Slide Time: 10:56)

Sample Correlation Coefficient

$$r_{xy} = (\text{sign of } b_1) \sqrt{\text{Coefficient of Determination}}$$

$$r_{xy} = (\text{sign of } b_1) \sqrt{r^2}$$

$$\hat{y} = b_0 + b_1 x$$

where:

b_1 = the slope of the estimated regression equation



From r square we have to find out the r that is a correlation coefficient. So, the sample correlation coefficient r_{xy} equal to sign of b_1 in our problem it is the sign of b_1 is positive root

of coefficient of determination r square so sign of b₁ into r square. So, what is this b₁ is that the slope of the regression equation.

(Refer Slide Time: 11:23)

Sample Correlation Coefficient

$$r_{xy} = (\text{sign of } b_1) \sqrt{r^2}$$

The sign of b_1 in the equation $\hat{y} = 10 + 5x$ is "+".

$$r_{xy} = +\sqrt{.8772}$$

$$r_{xy} = +.9366$$



In our problem it is y equal to $10 + 5 x$ so the sign is + the root of 0.8772 is this is your correlation coefficient and remember that the range of correlation coefficient is -1/2 to 1 but the range of r square is 0 to 1. Here in the correlation coefficient if it is -1 it is a perfectly negative correlation if it is +1 it is perfectly positive correlation if it is 0 there is no correlation. In the context of r square if it is 1 it is a perfect model that means all variability of y can be explained with the help of independent variable. If it is 0 there is no relation relationship between x and y .

(Refer Slide Time: 12:13)

Assumptions About the Error Term e

1. The error e is a random variable with mean of zero.
2. The variance of e , denoted by e^2 , is the same for all values of the independent variable.
3. The values of e are independent.
4. The error e is a normally distributed random variable.



Another important point is assumptions about the error term e , here the two tests the goodness of the model not only r^2 is important you have to plot the error term. When you look at the error term we have to look at the behavior of that error is nothing but actual minus predicted value, so what are the assumption of the error term the error ' e ' is he a random variable with mean equal to 0, so the error has to be appear in a random manner where the sum of positive error should be equal to sum of negative errors so that sum will be 0.

The variance of e denoted by e^2 is the same for all values of the independent variable, so that a concept called a homoscedasticity what is the meaning of that one is if there are many x_1 say x_2 x_3 independent variable these variance of x_1 variance of x_2 variance of x_3 should be same then only there is a meaningful comparison otherwise the variance of the error should be the same then only there is a meaningful comparison.

And the value of e is independent another important there should not be any pattern in the error term sometime what will happen when you plot the error term sometimes there is an increasing trend sometimes there may be in decreasing trend this kind of, this kind of pattern is not allowed the error term has to be distributed randomly. And the another point is the error e is normally distributed random variable now testing for significance.

(Refer Slide Time: 14:03)

Testing for Significance

- To test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of β_1 is zero.
- Two tests are commonly used:

t Test and **F Test**

- Both the t test and F test require an estimate of s^2 , the variance of e in the regression model.



12

So far as I told you in the beginning of the class whatever regression equations and the goodness of model which you have tested only for the sample data what is the sample data y equal to $b_0 + b_1 x$ so this capital Y equal to $\beta_0 + \beta_1 X$ whatever we have know done is only for the sample. Now we are going to see whether this model is valid even at the population level for that purpose we are going to do some assumption we will see what is that assumption that is a hypothesis?

To test for a significant regression relationship we must conduct a hypothesis test to determine whether the value of β_1 is 0. What will happen if the β_1 is 0 there is no relation between x and y at the population level. But there is a possibility that there may be a relation between x and y at the sample data it is not necessary that even at the population level there will be a relation between x and y . so, that testing can be done by two methods one is a t-test another one is F test.

Both t-test and F tests require an estimation of S^2 . S^2 is called the variance of the error otherwise if you say s it is the standard error the variance of e in the regression model.

(Refer Slide Time: 15:22)

Estimate of s

- An Estimate of s

The mean square error (MSE) provides the estimate of s^2 , and the notation s^2 is also used.

$$s^2 = \text{MSE} = \frac{\text{SSE}}{(n - 2)}$$

where:

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

What is the estimation of the standard error suppose in years normal data set if there are two data set data set 1 and 2, 1 is having lesser variance than the other one so the first data it is more homogeneous in the same way when you do in regression model suppose there are two model is there model 1 model 2 for the model in which there is a lesser standard error that is a more

suitable model. So, the mean square error provides the estimate of S Square and the notation S^2 is used so s square is nothing but MSE mean squared error.

We know that how we got to MSE, MSE is SSE a divided by $n - 2$ here what the degrees of freedom $n - 2$ so the logic is $n - 1 - K$ there is a logic of degrees of freedom. K is number of independent variable in this we are having only one independent variable we know that already the degrees of freedom is $n - 1$, so $n - 1 - K$ will be $n - 2$ and SSE also you see that we can find out that formula which I in the beginning of the class which I am saying $y_i - \hat{y}$ whole square. The \hat{y} you can substitute b_0 actually it is $b_0 + b_1 x$ when you bring - inside b_0 , $- b_0 - b_1 x$.

(Refer Slide Time: 16:51)

Testing for Significance

- An Estimate of s
 - To estimate s we take the square root of s^2 .
 - The resulting s is called the standard error of the estimate.

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}}$$



The term S we make the square root of s square the resulting is called standard at other term. So, when you take the square root of this mean squared error so you will get the standard error see in our problem I will go back I will see what is the standard error here where it is standard error as I told you by using shortcut method you can use SSE divided by $n - 2$ so that is MSE so $S_{xy} - S_{xy}$ S_{xy} whole square by S_{xx} divided by $n - 2$ that is the standard error of the estimate. So you have to take the square root of that then you look at the standard error.

(Refer Slide Time: 17:37)

Testing for Significance: t Test

- Hypotheses

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

- Test Statistic

$$t = \frac{b_1}{s_{b_1}}$$

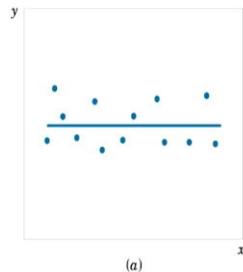


Now you go for hypothesis testing so what is the hypothesis testing beta 1 equal to 0 that means that there is no relation between x and y. Alternate hypothesis is beta 1 $\neq 0$ the test statistic is $b_1 - \beta_1$ divided by s_{b_1} . s_{b_1} is the standard error for the coefficient of b. So, since beta 1 we are assuming 0 it is simply b_1 divided by s_{b_1} .

(Refer Slide Time: 18:06)

Case 1

$$H_0: \beta_1 = 0$$



(a)

In this case hypothesis is not rejected

17

What is the meaning of b_1 beta 1 equal to 0 that means there is no relation between x and y. If it is when you plot the data in this case hypothesis is accepted because beta 1 equal to 0 now what is happening the beta 1 is not equal to 0 you see for this kind of data set so there is some relation between x and y see in this case the hypothesis is rejected so we are saying beta 1 not equal to 0.

(Refer Slide Time: 18:30)

The Standard Deviation of the Regression Slope

- The standard error of the regression slope coefficient (b_1) is estimated by

$$S_{b_1} = \frac{S_e}{\sqrt{\sum (x - \bar{x})^2}} = \frac{S_e}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

where:

S_{b_1} = Estimate of the standard error of the least squares slope

$S_e = \sqrt{\frac{SSE}{n-2}}$ = Sample standard error of the estimate



So, this is very important how to find out the standard error of the coefficient of x that is the b_1 , so S_{b_1} is S_e that is a standard error will divided by root of see $\sum (x - \bar{x})^2$, you see intuitively the total error that is S_e then we are dividing how much error is due to this independent variable, so total error did away portions of error from independent variable x. So, that will give you S_{b_1} .

(Refer Slide Time: 19:04)

Testing for Significance: t Test

■ Rejection Rule

Reject H_0 if $p\text{-value} \leq \alpha$
or $t \leq -t_{\alpha/2}$ or $t \geq t_{\alpha/2}$

where:

$t_{\alpha/2}$ is based on a t distribution
with $n - 2$ degrees of freedom



What is the rejection rule reject H_0 if the p-value is less than or equal to alpha we have seen many times this one so what will happen this one if the p value, the p value is the see alpha the p-value is less than that you are to reject it otherwise accept it, where $t_{\alpha/2}$ is the two-tailed test because we are writing beta equal to 0 Beta Beta 1 equal to 0, beta 0 equal to 0 and when you look at the t table you were to see $n - 2$ degrees of freedom.

(Refer Slide Time: 19:42)

Testing for Significance: t Test

1. Determine the hypotheses.

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

2. Specify the level of significance. $\alpha = .05$

3. Select the test statistic.

$$t = \frac{b_1}{S_{b_1}}$$

4. State the rejection rule.
Reject H_0 if $p\text{-value} \leq .05$
or $|t| > 3.182$ (with
3 degrees of freedom)

So, first determine the hypothesis beta 1 equal to 0 beta 1 not equal to 0 specify the significant level alpha equal to 5% select the test statistics b 1 by Sb1 state the rejection role reject H 0 if the p-value is less than or equal to 0.05 otherwise the t is greater than 3.182 when n - 2 degrees of freedom there are 5 data so 5 - 2 is the 3 degrees of freedom.

(Refer Slide Time: 20:16)

Testing for Significance: t Test

5. Compute the value of the test statistic.

$$t = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{5}{1.08} = 4.63$$

6. Determine whether to reject H_0 .

$t = 4.541$ provides an area of .01 in the upper tail. Hence, the $p\text{-value}$ is less than .02. (Also, $t = 4.63 > 3.182$.) We can reject H_0 .

So, this was the compute the value of the test statistics so 5 data by the Sb 1 is 1.08 we get 4.63 determine whether to reject H 0 because t equal to 4.5 form provides in the area of 0.01 in the upper tail here is the p-value is less then we will see how the got the p-value less than 0.02 so P

is greater than 4.63 we can reject null hypothesis and we reject null hypothesis what we are conclude there is a relation between x and y.

(Refer Slide Time: 20:51)

Hypothesis Tests for the Slope of the Regression Model

$H_0: \beta_1 = 0$	$t = \frac{b_1 - \beta_1}{S_b}$
$H_1: \beta_1 \neq 0$	where: $S_b = \frac{S_e}{\sqrt{SS_{xx}}}$
$H_0: \beta_1 \leq 0$	$S_e = \sqrt{\frac{SSE}{n-2}}$
$H_1: \beta_1 > 0$	$SS_{xx} = \sum_{i=1}^n X_i^2 - \frac{(\sum X_i)^2}{n}$
$H_0: \beta_1 \geq 0$	$\beta_1 = \text{the hypothesized slope}$
$H_1: \beta_1 < 0$	$df = n - 2$





This was the, here also $H_0: \beta_1 = 0$, $H_1: \beta_1 \neq 0$, 2 tail test, this is the right tailed test this is a left tailed test this was the formula for finding $t = (b_1 - \beta_1) / S_b$ to find $S_b = S_e$ is divided by root of SS_{xx} that is this is a form of $S_e = \sqrt{SSE / (n - 2)}$, $SS_{xx} = \sum (X_i^2 - (\sum X_i)^2 / n)$ remember it is $n - 2$ degrees of freedom.

(Refer Slide Time: 21:19)

Confidence Interval for β_1

- We can use a 95% confidence interval for β_1 to test the hypotheses just used in the t test.
- H_0 is rejected if the hypothesized value of β_1 is not included in the confidence interval for β_1 .





Next we can use the 95% confidence interval for beta 1 to test the hypothesis just used in the test. So, now with the help of conference interval also we can decide whether null hypothesis should

be accepted or rejected it is not as rejected if the hypothesis value of b_1 is not included in the confidence interval. our b_1 value what we are assuming to 0 so in that confidence interval if the 0 is appearing we have to accept the null hypothesis otherwise we have to reject null hypothesis.

(Refer Slide Time: 21:55)

Confidence Interval for β_1

- The form of a confidence interval for β_1 is:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

b_1 is the point estimator

$t_{\alpha/2} s_{b_1}$ is the margin of error

Where $t_{\alpha/2}$ is the t value providing an area of $\alpha/2$ in the upper tail of a t distribution with $n - 2$ degrees of freedom

So, confidence interval for beta 1 is the form of controlled $b_1 +$ or $- t \alpha/2 S_{b_1}$, b_1 which you got from our regression equation that is a coefficient of x_1 , S_{b_1} previously we are getting out I told you what is the formula for getting is b_1 , so when you substitute it here.

(Refer Slide Time: 22:22)

Confidence Interval for β_1

- Rejection Rule

Reject H_0 if 0 is not included in the confidence interval for β_1 .

- 95% Confidence Interval for β_1

$$b_1 \pm t_{\alpha/2} s_{b_1} = 5 \pm 3.182(1.08) = 5 \pm 3.44$$

or 1.56 to 8.44

- Conclusion

0 is not included in the confidence interval.

Reject H_0

So, what is getting b_1 is 5 + or - $t_{\alpha/2}$ is 3 point 1 8 - S_{b_1} is 1.08 so 5 + or - 3.44 so lower limit is 1.56 upper limit is 8.44 you see in that there is no 0's there so we have to reject our null

hypothesis, conclusion 0 is not included in the confidence interval so we are rejecting null hypothesis.

(Refer Slide Time: 22:43)

Testing for Significance: F Test

- Hypotheses
$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$
- Test Statistic
$$F = MSR/MSE$$



The previous way we have used the t-test some time what will happen if the number of independent variable is more than 2 we have to do the t-test to 2 times. If there are say 5 independent variable you have to do file individually as I told you whenever you are comparing more than two we should go for Anova that is the F-test so here also whenever there is a number of independent variables more otherwise a generic method for testing the beta 1 equal to 0 hypothesis is going for F test.

So here you have F test is a MSR divided by MSE even in Anova also you know anova what we write is we write MS treatment divided by MSE, MS treatment is nothing but our regression sum of square mean regression sum of square.

(Refer Slide Time: 23:43)

F-Test for Significance

- F Test statistic:

$$F = \frac{MSR}{MSE}$$

where

$$MSR = \frac{SSR}{k}$$
$$MSE = \frac{SSE}{n - k - 1}$$

where F follows an F distribution with k numerator degrees of freedom
and (n - k - 1) denominator degrees of freedom
(k = the number of independent variables in the regression model)



So, F equal to MSR divided by MSE you see that MSR how we are getting MSR SSR divided by K and K is number of degrees of freedom that is nothing but number of independent variable n - K - 1 is degrees of freedom for the error term.

(Refer Slide Time: 24:02)

Testing for Significance: F Test

- Rejection Rule

Reject H_0 if
 $p\text{-value} \leq \alpha$
or $F \geq F_\alpha$

where:

F_α is based on an F distribution with
1 degree of freedom in the numerator and
 $n - 2$ degrees of freedom in the denominator



So, what is the rejection rule reject H_0 if the p-value is less than equal to alpha otherwise if the calculated F value is greater than the value which we got from the table. So, if alpha is based on the F distribution we have to look at the what is the degrees of freedom as you look at you see enumerated degrees of freedom in this problem we have only one independent variables so one degrees of freedom numerator so $n - 2$ is $5 - 2 = 3$, degrees of freedom for denominator.

(Refer Slide Time: 24:33)

Testing for Significance: F Test

1. Determine the hypotheses.

$$H_0: \beta_1 = 0$$

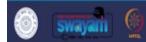
$$H_a: \beta_1 \neq 0$$

2. Specify the level of significance. $\alpha = .05$

3. Select the test statistic. $F = \frac{MSR}{MSE}$

4. State the rejection rule.

Reject H_0 if $p\text{-value} \leq .05$
or $F \geq 10.13$ (with 1 d.f.
in numerator and
3 d.f. in denominator)



So, beta 1 equal to 0 beta 0 equal to 0 alpha equal to 0.05 F equal to MSR divided by MSE, so p value we got to find out so numerator degrees of freedom is 1 t nominated is a freedom is 3 so that p value is 10.13.

(Refer Slide Time: 24:51)

Jupyter Code

```
In [2]: import numpy as np
import matplotlib.pyplot as plt

In [3]: import seaborn as sns

In [4]: import pandas as pd
import matplotlib as mpl
import statsmodels.formula.api as sm
from sklearn.linear_model import LinearRegression
from scipy import stats

In [5]: tb1 = pd.read_excel('C:/Users/Somi/Documents/regr.xlsx')
```



31

So, now we will use Python code rules that will do that import numpy as np import matplotlib.pyplot as plt, import seaborn ssn, import pandas as pd input matplotlib as mpl, import stats.models.formula.api as sm, from sklearn.linear_model import LinearRegression, from scipy import stats. So, tb1 we are going to that regression data we are going to save in the object called tb1 and we are reading that.

Now what is happening you see that the p-value for the TV ad we say alpha equal to 5% it is less than 0.01 so TV ad is insignificant variable if it is more than 5% say, if it is a 0.06 the regression equation we will not include this independent variable TV ads you have statistics 21.43 I will go back will verify this answer that we can find out MSR, how we do MSR and that is what I am saying that time the first you have to find out SSR regression sum of square regression sum of square is see $\Sigma (y\hat{ } - \bar{y})^2$ divided by k, k is number of independent variable will get MSR.

So the p-value sorry the F value is 21.43 and see the probability it is less than 0.01 so that is less than 0.05 so we are saying that the model as a whole there are two things is there as a whole model the F value is less than 0.05 the model is valid and if you want to check individual independent variable also. So, see here it is less than 0.05 so this variable is significant see the lower limit, upper limit there is no 0 here 1.563, 8.43. So we can we cannot accept where to reject null hypothesis.

(Refer Slide Time: 27:02)

Testing for Significance: F Test

5. Compute the value of the test statistic.

$$F = \text{MSR}/\text{MSE} = 100/4.667 = 21.43$$

6. Determine whether to reject H_0 .

$F = 17.44$ provides an area of .025 in the upper tail. Thus, the p -value corresponding to $F = 21.43$ is less than $2(.025) = .05$. Hence, we reject H_0 .

The statistical evidence is sufficient to conclude that we have a significant relationship between the number of TV ads aired and the number of cars sold.



You see MSR is 21.43 we are here 21.43 so we can verify that our Python result then what we have done it with the help of manually. Some cautions about the interpretation of significant test is so it is very important this one rejecting H_0 : b_1 or beta 1 equal to 0 and concluding that the relationship between x and y significant does not enable us to conclude that there is a cause and effect relationship in present between x and y. Just because of there is a correlation we cannot say there is a cause-and-effect relationship.

So, just because of we are able to reject H_0 : $\beta_1 = 0$ and demonstrate statistical significant does not enable us to conclude that there is a linear relation between x and y . Dear students in this class what we have seen we have taken one sample problem then we have fitted a regression equation in the regression equation we gone for hypothesis testing we have tested the significance of that independent variables.

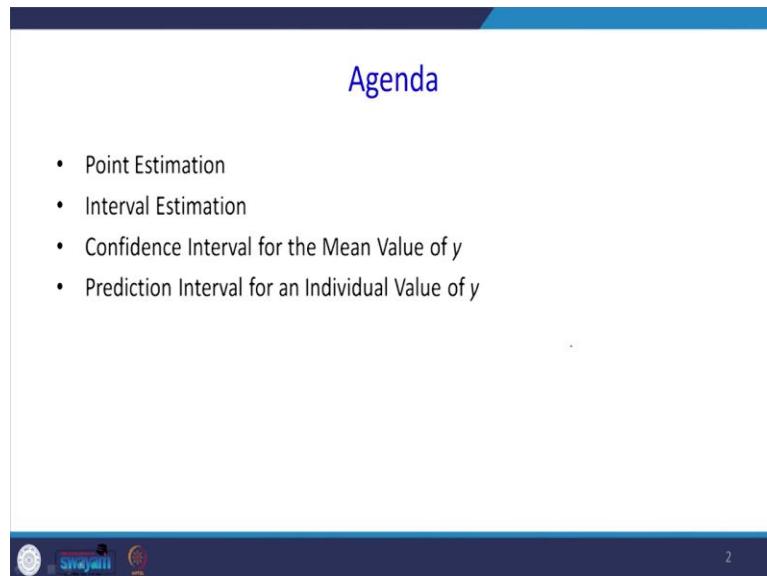
There are two way to do the significance test one is by using t method t statistic method another one is F test method in both method we all got the same answer. Then I have explained what is the meaning of coefficient of determination, that is r^2 from the r square I have explained how we can get the r. The next class will go for multiple regression equation where we will consider more than one independent variable and we will also ill explain some important assumptions in the regression equations. Thank you very much

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 31
Estimation Prediction of Regression Model Residual Analysis: Validating Model
Assumptions - 1

Dear students in the previous class we have explained I have explained the confidence interval for the x coefficient that is the b. For the b we have found what was the lower limit and upper limit. In this class we will find the confidence interval for y and prediction interval.

(Refer Slide Time: 00:43)



Agenda

- Point Estimation
- Interval Estimation
- Confidence Interval for the Mean Value of y
- Prediction Interval for an Individual Value of y

2

So, today's class agenda is we list will explain what is the point estimate and interval estimate; and confidence interval for the mean value of y and prediction interval for the individual value of y .

(Refer Slide Time: 00:43)

Problem

- Data were collected from a sample of 10 ice cream vendors located near college campuses.
- For the i^{th} observation or restaurant in the sample, x_i is the size of the student population (in thousands) and y_i is the quarterly sales (in thousands of dollars).
- The values of x_i and y_i for the 10 restaurants in the sample are summarized in Table

3

We will take one problem then I first I will solve this problem with the help of Python then I will explain what is the meaning of this confidence interval and prediction interval. Data were collected from a sample of 10 ice cream vendors located near college campuses for the i^{th} observation our restaurant in the sample x_i is the size of the student population and y is the quarterly sales of ice cream. The values of x_i and y_i for 10 restaurants in the sample are summarized in the table this is given x 1.

(Refer Slide Time: 01:29)

Data

Restaurant	Student Population (1000)	Sales (1000)
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

4

So, what the problems is the independent variable is student population dependent variable is sales there was a 10 data set like this.

(Refer Slide Time: 01:40)

Python code for scatter plot

```
In [4]: import pandas as pd
import matplotlib as mpl
import statsmodels.formula.api as sm
from sklearn.linear_model import LinearRegression
from scipy import stats
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt

In [5]: data = pd.read_excel('C:/Users/Somi/Documents/lrm.xlsx')
data
```

	Restaurant	Student Population	Sales
0	1	2	58
1	2	6	105
2	3	8	88
3	4	8	118
4	5	12	117
5	6	16	137
6	7	20	157
7	8	20	169
8	9	22	149
9	10	26	202

5

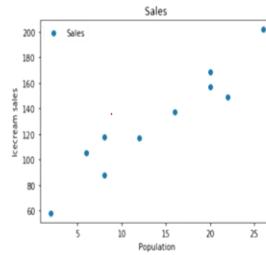
For given data set first we will run the regression model so import pandas pd import matplotlib as mpl import stats model dot formula dot api as sm from sklearn dot linear underscore model input linear regression from scipy import stats import Seaborn as sns, import numpy as np import matplotlib dot pyplot as plt. First we load the data we will treat the data there is a pd dot breed underscore this is a path where I have stored my excel file, so this is the data.

So, what is it there is a 10 dataset 10 restaurants this is a student population this student populations in terms of 1000 sales also in terms of 1000 of it, for a product called ice cream first we will plot the scatter plot.

(Refer Slide Time: 02:29)

Python code for scatter plot

```
In [36]: data.plot('Population', 'Sales', style='o')
plt.ylabel('Icecream sales')
plt.title('Sales ')
plt.show()
```



6

So, data dot plot population, sales style equal to 'o', so why label is ice cream sales title is sales when we they show this versus, so what is happening there seems to be some positive trend when the student population is more there is a more number of sales. We will find a regression model for this.

(Refer Slide Time: 02:56)

```
In [17]: import statsmodels.api as s
St_pop = data['Population']
sales = data['Sales']
St_pop = sm.add_constant(St_pop)
model1 = sm.OLS(sales, St_pop)
results1 = model1.fit()
print(results1.summary())

OLS Regression Results
==============================================================================
Dep. Variable: Sales   R-squared:      0.903
Model: OLS            Adj. R-squared:  0.891
Method: Least Squares F-statistic:   74.25
Date:    Wed, 04 Sep 2019   Log-likelihood: -19.342
Time:    14:33:11   Log-likelihood: -19.342
No. Observations: 10   AIC:         82.68
Df Residuals:     8   BIC:         83.29
Df Model:        1
Covariance Type: nonrobust
==============================================================================
              coef    std err          t      P>|t|      [0.025  0.975]
-----+-----+-----+-----+-----+-----+-----+
const   60.0000    0.200    300.000    0.000   34.725   85.275
Population  5.0000    0.580     8.517    0.000    3.462    6.338
-----+-----+-----+-----+-----+-----+-----+
Omnibus: 0.928 Durbin-Watson: 3.224
Prob(Omnibus): 0.629 Jarque-Bera (JB): 0.616
Skew: -0.008 Prob(JB): 0.735
Kurtosis: 1.790 Cond. No. 33.6
=====+=====+=====+=====+=====+=====+=====+
```

So import startsmodels dot api as s, St_pop equal to that is a student population equal to data I am going to in the population I am going to stay a store variable called St_pop, sales equal to data sales St_pop = s.add_constant because I need to have the constant in the regression equation. So, model one equal to sm.OLS(sales, sale is our dependent variable St_pop) is our independent variable result 1 equal to model one dot fit. So, print result one dot summary.

So what we are getting here and you look at this, this is the constant value. So, the sales equal to I can write sales equal to 60+5 this is our independent variable say population st underscore. It is a population. So, what is the meaning of interpretation of this file if the student population is increased by one unit the sales will increased by 5 units look at this R square R square is very good that is 90.3 I will explain meaning of adjusted R square in coming class.

Then we have to remember this is there is a standard at it is 0.58 the t value is 8.68 this is a problem the probability is 0.00 this is lower limit this is upper limit. There is another way we can write it otherwise directly we can get y-intercept and x coefficient from a sklearn dot linear

underscore model input linear regression x equal to data ['population'] dot values reshaped (-1, 1), y equal to data['sales'] dot value.reshape (-1, 1), reg = LinearRegression(), reg.fit is x, y so linear regression is copy underscore x equal to true fit underscore intercept true, equal to true, n underscore jobs equal to one normalize equal to false.

What is the meaning of fit underscore intercept sometime if you put false so sometimes when you fit a regression line suppose it is coming like this, so there is a y intercept is a this much distance is y intercept. Sometime you need not have the y intercept for that time for that time you have to use write false. The another one is normalized equal to false so there are y and there are x value you if you normalize x-value and y-value then you run the regression they will get a standardized regression coefficient.

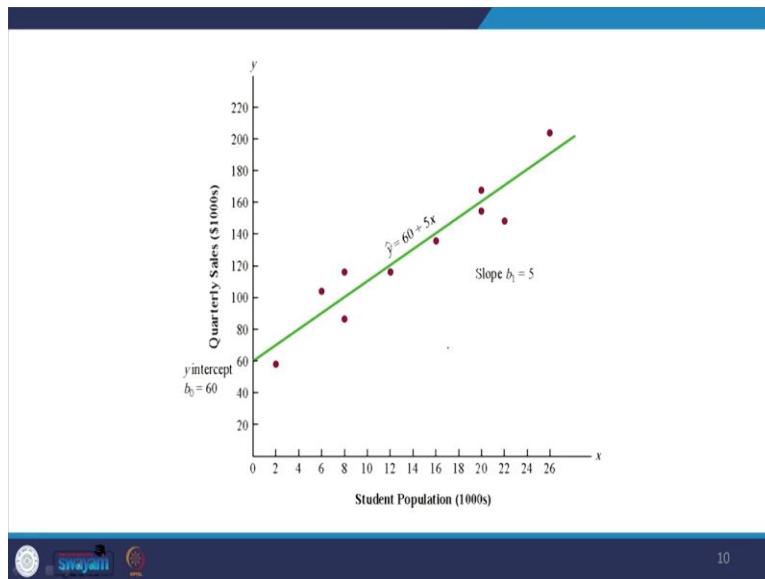
Now we are not equal to false is written so we are not going to normalize the data set so reg dot intercept underscore 0, reg coefficient underscores 0, 0 so this is your intercept this is your x coefficient. The previous also you look at the previous slide there also here got the 60 and 5 same result.

(Refer Slide Time: 06:08)

The slide has a blue header bar with the title 'Python code for regression'. The main content area contains a bulleted list and a mathematical equation. The list states: 'In the Ice cream vendor example, the estimated regression equation $60 + 5x$ provides an estimate of the relationship between the size of the student population x and quarterly sales y '. Below the list is the equation $\hat{y} = 60 + 5x$. At the bottom of the slide, there is a dark footer bar with three small icons on the left and the number '9' on the right.

So, what we can do in the ice cream at our example the estimated regression equation is $60 + 5x$ provides an estimate of the relationship between the size of the student population x and quarterly sales y . so, this is our regression equation.

(Refer Slide Time: 06:27)



So, this is our regression equation so the y intercept the slope is 5 the y intercept is 60.

(Refer Slide Time: 06:38)

Point Estimate

- We can use the estimated regression equation to develop a point estimate of the mean value of y for a particular value of x or to predict an individual value of y corresponding to a given value of x .
- For instance, suppose a manager want a point estimate of the mean quarterly sales for all restaurants located near college campuses with 10,000 students.

Then we will see what is the point estimate, we can use the estimated regression equation to develop your point estimate of the mean value of y for a particular value of x or to predict an individual value of y corresponding to a given value of x . So, whatever value which you are predicting is the mean value there is another we can predict an individual value. For instance suppose your manager want to want see a point estimate of the mean quarterly sales for all restaurants here you have to see all restaurants located nearby a college campus with the 10,000 students.

(Refer Slide Time: 07:23)

Point estimate

- Using the estimated regression equation $60 + 5x$, we see that for $x = 10$ (or 10,000 students), $60 + 5(10) = 110$.
- Thus, a point estimate of the mean quarterly sales for all restaurants located near campuses with 10,000 students is \$110,000.

```
In [32]: reg.predict(10)
Out[32]: array([[110.]])
```

12

So if you say student population is 10 what will happen when you substitute to 10 it is 110. So, that is your point estimate for the mean quarterly sales of all restaurant located near campus is the 10,000 students a one lakh 10,000 dollar. So, even regression equation also we can use the predict function `reg.predict` when you put the input value that is x value you can get y value is 110.

(Refer Slide Time: 07:45)

Point estimate

- Now suppose the manager want to predict sales for **an individual restaurant** located near College, with 10,000 students.
- In this case we **are not interested in the mean value** for all restaurants located near campuses with 10,000 students;
- We are just interested in predicting quarterly sales for **one individual restaurant**.
- As it turns out, the point estimate for an **individual value** of y is the same as the point estimate for the mean value of y .
- Hence, we would predict quarterly sales of $60 + 5(10) = 110$ or \$110,000 for this one restaurant.

13

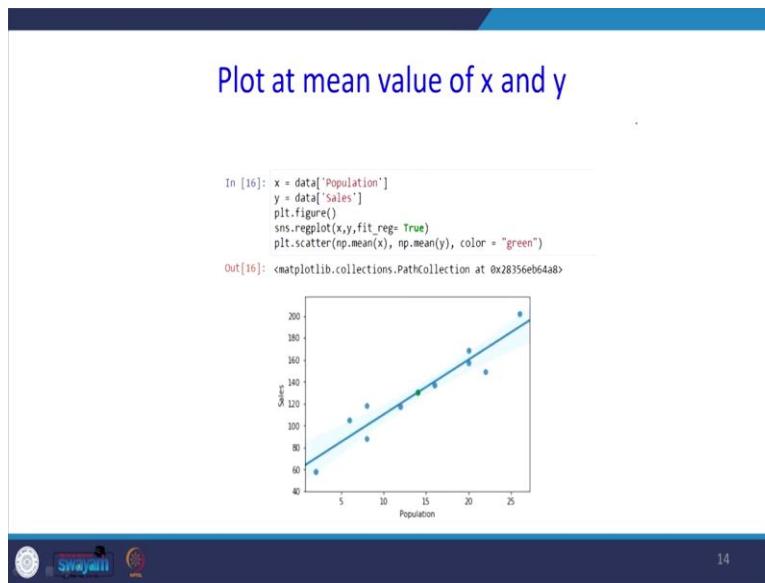
So, what is a point estimate now suppose the manager want to predict the sales of an individual restaurant located in area college with the 10,000 students. In this case we are not interested in the mean value of all restaurants located near compass of the 10,000 students we are just

interested predicting quarterly sales of one individual restaurant as it turns out the point estimate for an individual value of y is the same as the point estimate for the mean value of y .

Hence we would predict quarterly sells sales $60 + 5$ the 10 is our input it is 110000 dollars so what I am saying for you find estimate the value of confidence interval and the value of prediction interval is same you see that you may see the similarity also here and when you go for hypothesis testing see $x\bar{ } +$ or $- Z \Sigma$ by root n, right. So, this $x\bar{ }$ is nothing but our point estimate so whatever value after substituting 10 we are getting 110 we are getting that is only point estimate so point estimate is not the reliable one so we need to have interval estimate.

So, interval estimate in the hypothesis testing context $x\bar{ } + Z \Sigma$ by root n is the upper limit $x\bar{ } - Z \Sigma$ by root n lower limit. How we are going to find out upper limit lower limit in the regression context I will explain in the next slide.

(Refer Slide Time: 09:17)



First we will plot it so what will happen plot at mean value of x and y so x equal to data population y equal to data sales so x is the population y is the sales value plot dot figure, sns dot reg plot x, y fit underscore regression is true plot dot scatter np dot mean value of x , np dot mean value of y so we got this regression equation. You see that if you want to draw the best regression equation that has to pass through $x\bar{ }$, $y\bar{ }$. So, that is why, so this point is mean of x so this point is mean of y .

(Refer Slide Time: 10:06)

Confidence Interval Estimation

- **Confidence interval**, is an interval estimate of the *mean value of y* for a given value of x.
- **Prediction interval**, is used whenever we want an interval estimate of an *individual value of y* for a given value of x.
- The point estimate of the *mean value of y* is the same as the point estimate of an *individual value* of y.
- The margin of error is larger for a prediction interval.

15

So what is a confidence interval estimation, confidence interval is an interval estimate for the mean value of y for a given value of x. But the prediction interval is used whenever we want an interval estimate of an individual value of y right this is an individual value of y that is a mean value of y for a given value of x. For example y it is a mean value so what we are predicting is expected value of $E(y) = a + bx$ so whatever value after substituting x we are getting into the mean value.

So what will happen the margin of error is larger for your prediction interval. So, the prediction interval the margin of error will be larger for your confidence interval the margin of error will be smaller.

(Refer Slide Time: 10:56)

Confidence Interval Estimation

x_p = the particular or given value of the independent variable x

y_p = the value of the dependent variable y corresponding to the given x_p

$E(y_p)$ = the mean or expected value of the dependent variable 'y' corresponding to the given x_p

$\hat{y} = b_0 + b_1 x_p$ = the point estimate of $E(y_p)$ when $x = x_p$

$$60 + 5(10) = 110.$$

16

So, confidence interval of estimation, for example take x_p equal to the particular or given value of independent variable x, y_p is the value of the dependent variable y corresponding to the given x_p , so expected value of y_p is nothing but mean or expected value of dependent variable y corresponding to the given x_p , so \hat{y} equal to $b_0 + \beta_1 x_p$, is the point estimate of expected value of y_p when x equal to x_p so that is why $60 + 5*10$ is 110.

(Refer Slide Time: 11:47)

Confidence Interval Estimation

In general, we cannot expect \hat{y}_p to equal $E(y_p)$ exactly.

If we want to make an inference about how close \hat{y}_p is to the true mean value $E(y_p)$, we will have to estimate the variance of \hat{y}_p .

The formula for estimating the variance of \hat{y}_p given x_p , denoted by $s_{\hat{y}_p}^2$, is

$$s_{\hat{y}_p}^2 = s^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

17

In general we cannot expect \hat{y}_p is equal to expected value of y_p exactly if you want to make an inference about how close \hat{y}_p is to the true mean value of expected value of y_p we will have to estimate the variance of \hat{y}_p . The formula for estimating the variance of \hat{y}_p at

given x_p is denoted by s^2 , \hat{y}_p so this \hat{y}_p is nothing but $s^2 ((1/n) + ((xp - \bar{x})^2 / \sum(x_i - \bar{x})^2))$, this is the variance of predicted y .

(Refer Slide Time: 12:36)

Confidence Interval Estimation

CONFIDENCE INTERVAL FOR $E(\hat{y}_p)$

$$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$$

$$s_{\hat{y}_p}^2 = s^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right]$$

$$s_{\hat{y}_p} = 13.829 \sqrt{\frac{1}{10} + \frac{(10 - 14)^2}{568}}$$

$$= 13.829 \sqrt{.1282} = 4.95$$

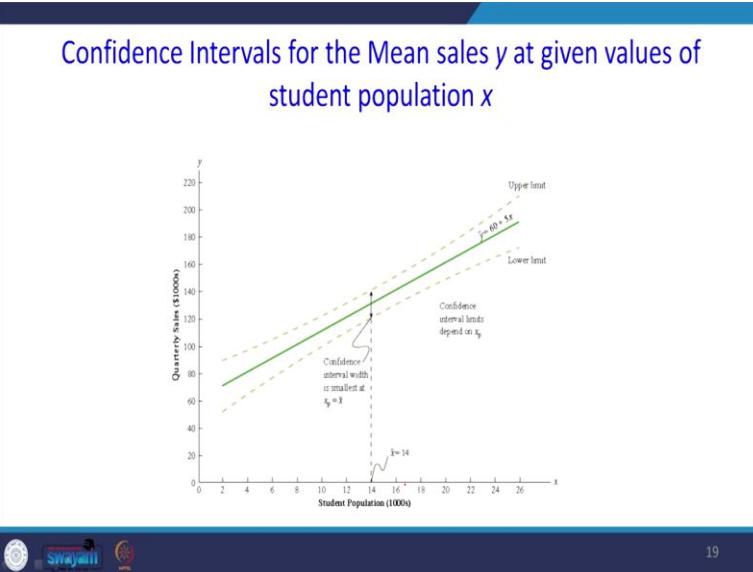
$$110 \pm 11.415$$

18

So, the confidence interval is you see that the confidence interval is we are writing \hat{y}_p + or - so this yes \hat{y}_p is the variance of this y at p right so what you are done previously in the hypothesis testing example x + or - z Sigma by root n . So, instead of \bar{x} we are writing \hat{y}_p + or - so instead of z we are writing $t_{\alpha/2}$ this standard error and so write we are writing $S_{\hat{y}_p}$, so that was the formula is equal to $s^2 (1/n)$ this can be derived a very easily.

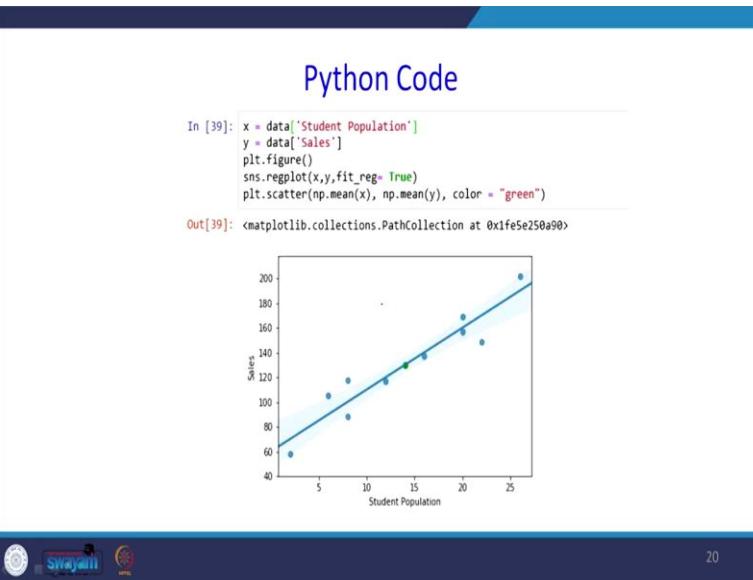
So I am not deriving you can refer any book for this. so, the variance of \hat{y}_p is equal to $s^2 ((1/n) + ((xp - \bar{x})^2 / \sum(x_i - \bar{x})^2))$, We can substitute this value here the s^2 is nothing but the standard error. So, we can substitute s^2 value n is 10 x_p is 10 because that is a value of x so \bar{x} is 14 whole square when you substitute 2 you are getting 110 + or - 11.415.

(Refer Slide Time: 13:44)



So, the Green Line shows green dotted line shows the upper limit the down one is shows the lower limit. You see that it is the confidence interval is not a straight line it is somewhat curved one so what is happening when x bar equal to 4 the interval now it is a very narrow. What will happen that is a special case.

(Refer Slide Time: 14:09)



Now we will plot this confidence interval okay what is happening here you see that when this Point, see it is not the straight line it is somewhat curved one. The confidence interal is very narrow when there is x equal to x bar we will see that a special case.

(Refer Slide Time: 14:27)

Special Case

The estimated standard deviation of \hat{y}_p

is smallest when $x_p = \bar{x}$ and the quantity $x_p - \bar{x} = 0$

$$s_{\hat{y}_p} = s \sqrt{\frac{1}{n} + \frac{(\bar{x} - \hat{x})^2}{\sum(x_i - \hat{x})^2}} = s \sqrt{\frac{1}{n}}$$

21

The estimated standard deviation of \hat{y}_p is smallest when x_p equal to \bar{x} . So, what will happen in the previous equation when you substitute x_p equal to \bar{x} so this term will become 0. So, remaining is s divided by $1/n$ you see that this is similar to our the result of central limit theorem. The variance of a sampling distribution is σ^2 by n it is similar to that.

(Refer Slide Time: 14:58)

Prediction Interval for an Individual Value of y

- Instead of estimating the mean value of sales for all restaurants located near campuses with 10,000 students, we want to estimate the sales for an individual restaurant located near a particular College with 10,000 students.
 - (1) The variance of individual ' y ' values about the mean $E(y_p)$, an estimate of which is given by s^2
 - (2) The variance associated with using \hat{y}_p estimate $E(\hat{y}_p)$, an estimate of which is given by $s_{\hat{y}_p}^2$

22

Now we will go for prediction interval for an individual value of y instead of estimating the mean value of sales for all restaurants located near campus of the trend of students we want to estimate sales on individual restaurant located near a particular college with the 10,000 students. So, when you go for predicting y value for an individual restaurant there are two component of variance has to be added one component is the variance of individual y values about the mean

value of y_p that is given by s^2 square, the variance associated with using \hat{y}_p estimate is expected value of y_p and estimate of which is given by s^2 square of \hat{y}_p .

So what is happening here if you want to go for a prediction interval these two variances has to be added one variances for y another variances for \hat{y}_p right.

(Refer Slide Time: 16:03)

Prediction Interval for an Individual Value of y

$$\begin{aligned}s_{\text{ind}}^2 &= s^2 + s_{\hat{y}_p}^2 \\&= s^2 + s^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right] \\&= s^2 \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right]\end{aligned}$$

23

You see that so that is where s^2 individual is s^2 square + s^2 \hat{y}_p so when you add it the s^2 square is common so we will get this formula. So, for this formula we will substitute the value when you substitute it you see it is a 14.69.

(Refer Slide Time: 16:17)

Prediction Interval for an Individual Value of y

$$\hat{y}_p \pm t_{\alpha/2} s_{\text{ind}}$$

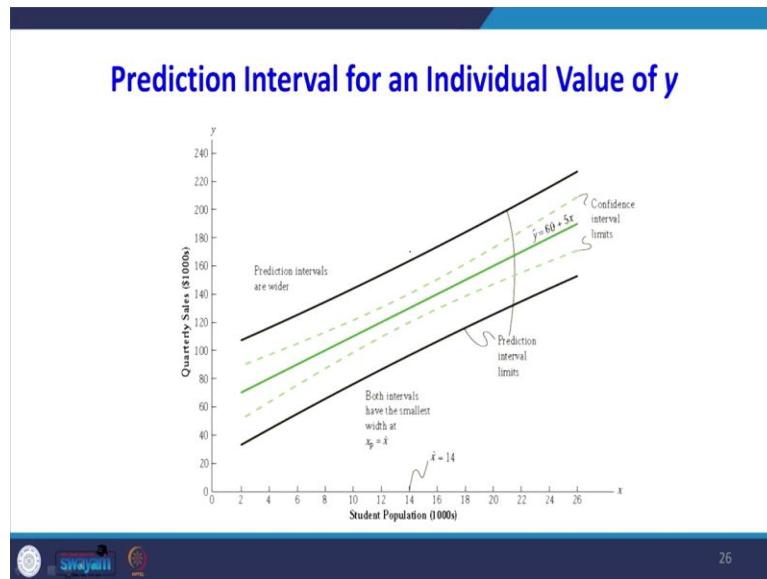
$$t_{\alpha/2} s_{\text{ind}} = 2.306(14.69) = 33.875,$$

$$110 \pm 33.875$$

25

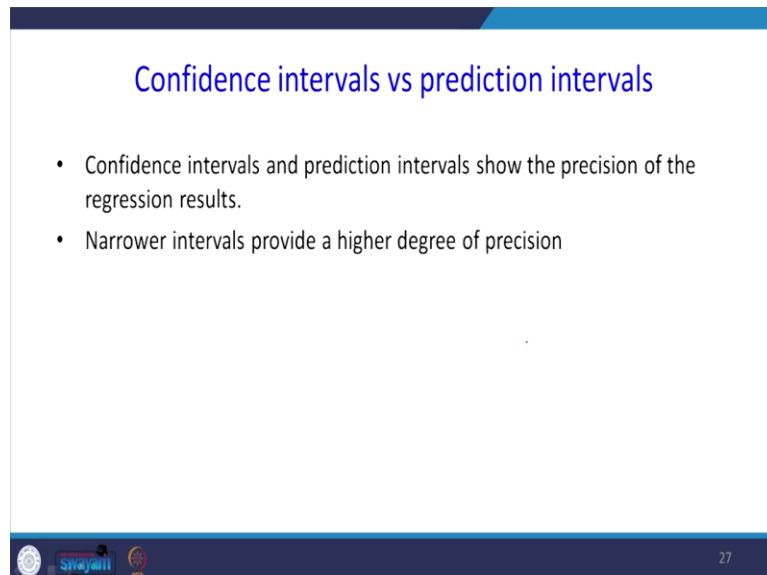
Now the value of t alpha by two and you look at the then when you substitute 14.69 you will get this was the this 33.875 is the margin of error so $110 +$ or $- 33.875$ will get so this one.

(Refer Slide Time: 16:37)



So, in the black line shows the prediction interval the Green Line shows the confidence interval both are not the straight line. So, when you look at this one see the prediction line is having margin of error is more compared to the confidence interval.

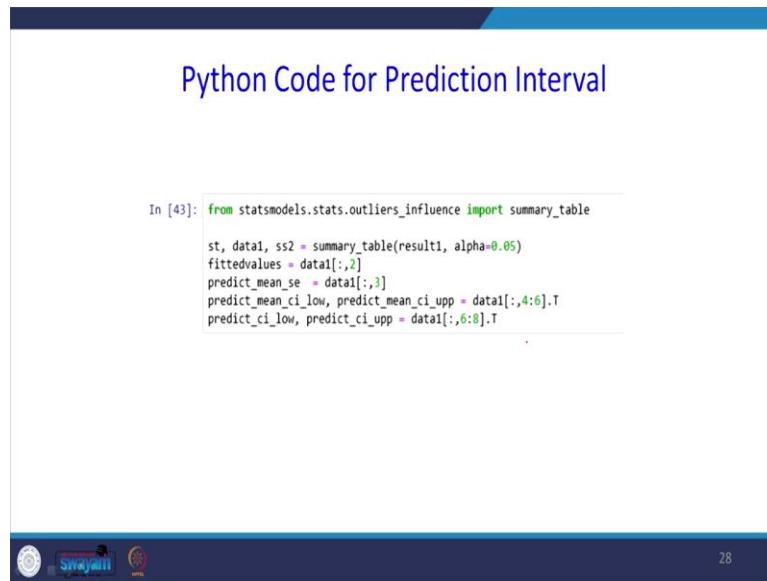
(Refer Slide Time: 16:58)



Now confidence interval versus prediction interval confidence intervals and prediction intervals show the precision of the regression result narrower intervals provide a higher degrees of

precision. So, what after doing regression analysis when you plot the confidence and prediction interval it has to be narrow if it is wide means that model is not the good model.

(Refer Slide Time: 17:23)



Python Code for Prediction Interval

```
In [43]: from statsmodels.stats.outliers_influence import summary_table
st, data1, ss2 = summary_table(result1, alpha=0.05)
fittedvalues = data1[:,2]
predict_mean_se = data1[:,3]
predict_mean_ci_low, predict_mean_ci_upp = data1[:,4:6].T
predict_ci_low, predict_ci_upp = data1[:,6:8].T
```

Now we will use Python to plot this prediction interval and confidence interval. So, for that purpose from statsmodels dot stats dot out layer underscore influence import summary table st command data 1, ss 2 equal to summary underscore table result 1, alpha equal to 5% fitted value is equal to data colon, second that means we are referring the third column predict underscore mean underscore ac equal to date data 1 colon, 3 that is we are referring fourth column predictor underscore mean ci interval that is mean ci interval means your confidence interval lower limit confidence interval upper limit.

So, that was because in their summary table that is in the summary table we are referring the fourth to sixth column, dot t predict underscore see a confidence table low-protein predict underscore ci upper limit data to 6 : 8. Actually what do you have is what is happening here we are getting in the summary table all the result so we are calling 4 to 6, 6 to 8, 3 2 to get a particular value that is the reason here.

(Refer Slide Time: 18:41)

Python Code

```
In [44]: predict_mean_ci_low  
Out[44]: array([ 51.02868339,  75.2931351 ,  87.10977127,  87.10977127,  
    109.56629088, 129.56629088, 147.10977127, 147.10977127,  
    155.2931351 , 171.03868339])  
  
In [45]: predict_mean_ci_upp  
Out[45]: array([ 88.96131661, 104.7060649 , 112.89022873, 112.89022873,  
    130.43781932, 150.43370192, 172.89022873, 172.89022873,  
    184.7068649 , 208.96131661])  
  
In [46]: predict_ci_low  
Out[46]: array([ 32.89834155,  54.8817226 ,  65.60291394,  65.60291394,  
    86.446108 , 106.446108 , 125.60291394, 125.60291394,  
    134.8817226 , 152.89834155])  
  
In [47]: predict_ci_upp  
Out[47]: array([107.10165845, 125.1182774 , 134.39708666, 134.39708666,  
    153.553892 , 173.553892 , 194.39708666, 194.39708666,  
    205.1182774 , 227.10165845])
```



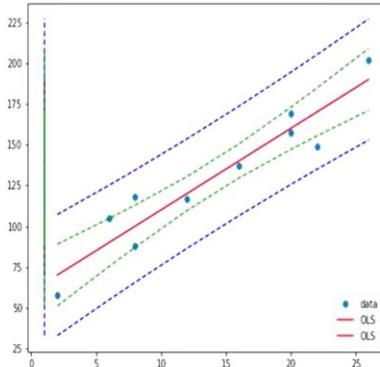
29

You see that this is the predict underscore main underscore ci_low so we are getting the confidence interval for the lower limit here predict mean underscore ci for upper limit. So, this was predict for this is a prediction interval this is for the confidence interval the first 2 things for the confidence interval the next bottom 2 is for the prediction interval. So, lower limit upper limit this is the lower limit see a ci_low, ci_upper is the upper limit.

(Refer Slide Time: 19:19)

Python Code

```
In [48]: X = s.add_constant(x)  
fig, ax = plt.subplots(figsize=(8,6))  
ax.plot(x, y, 'o', label="data")  
ax.plot(X, fittedvalues, 'r-', label="OLS")  
ax.plot(X, predict_ci_low, 'b--')  
ax.plot(X, predict_ci_upp, 'b--')  
ax.plot(X, predict_mean_ci_low, 'g--')  
ax.plot(X, predict_mean_ci_upp, 'g--')  
ax.legend(loc='best');  
plt.show()
```



30

So, this picture shows you see that x equal to s dot ad underscore constant fig, ax equal to plot dot subplot fig size equal to 8, 6, ax dot plot x, y ou label equal to data ax dot plot x, featured values are hypen, label Wireless ax plot dot x, predict underscore ci low it is in the dotted line by ax dot plot x, predict underscore ci underscore upper limit b hypen hypen ax dot plot x, predict

underscore mean ci low g a x dot plot x predict underscore main ci upper limit so the location is the best plt dot show.

So, when you run this command you will get this kind of here model. So, the green one shows the confidence level the blue one shows the prediction interval. In this picture when you look at the ‘r-‘ represents the red color b represents blue color g represents green color the hyphen represents what kind of pattern we need to have in the in the picture. Now what we have done in this class we have explained what is point interval and what is confidence interval and what is prediction interval.

So the point interval is same for particular value of x for both confidence and prediction interval. So, the another point which you have learnt in this lecture is that the confidence interval is not the straight line it is curved line similarly the prediction interval. The another one is the prediction interval is having more margin of error when compared to confidence interval after that what you have done with help of Python I have run this code to show you how to plot this confidence and prediction interval, thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 32
Estimation Prediction of Regression Model Residual Analysis: Validating Model
Assumptions - II

(Refer Slide Time: 00:30)

The slide has a dark blue header bar. The main title 'Lecture Objectives' is centered in blue text. Below it is a bulleted list of objectives:

- Understanding different types of residual analysis
- Plotting residual plots using python

In the bottom left corner, there are three small icons: a clock, a person, and a gear. In the bottom right corner, the number '2' is displayed.

This lecture is validating regression model assumptions. The lecture objective is understanding different types of residual analysis and plotting residual plots using Python.

(Refer Slide Time: 00:37)

The slide has a dark blue header bar. The main title 'Residual Analysis: Validating Model Assumptions' is centered in blue text. Below it is a bulleted list of points:

- *Residual analysis is the primary tool for determining whether the assumed regression model is appropriate*

Below the list, the text 'RESIDUAL FOR OBSERVATION i ' is followed by the formula $y_i - \hat{y}_i$. The word 'where' appears below the formula. To the right, two definitions are given:
 y_i is the observed value of the dependent variable
 \hat{y}_i is the estimated value of the dependent variable

In the bottom left corner, there are three small icons: a clock, a person, and a gear. In the bottom right corner, the number '3' is displayed.

Residual analysis validating model assumptions, first we will see what is the residual analysis. The residual analysis is the primary tool for determining whether the assumed regression model is appropriate. So, the residual for observation i is nothing but y_i there is an actual value and \hat{y}_i our predicted model. So, the difference between actual and predicted model it's nothing but the error otherwise you can call it is residual analysis.

So y_i is the observed value of dependent variable y \hat{y}_i is the estimated value of dependent variable.

(Refer Slide Time: 01:12)

The slide has a blue header bar with the text 'Assumptions about the error term .ξ'. Below the header is a white content area containing a mathematical equation and a list of four assumptions. At the bottom of the slide is a dark footer bar with three small icons on the left and the number '4' on the right.

$y = \beta_0 + \beta_1 x + \epsilon$

1. $E(\epsilon) = 0$.
2. The variance of ϵ , denoted by σ^2 , is the same for all values of x .
3. The values of ϵ are independent.
4. The error term ϵ has a normal distribution.

Assumptions about the error term that is Epsilon we know that y equal to $\beta_0 + \beta_1 x + \epsilon$. What are the assumption about this error term, number one the expected value of error is 0. The variance of error term denoted by Sigma square is the same for all values of x the values of error are independent the error term epsilon has their normal distribution. We will validate we will check this assumptions by drawing various residual plots in this lecture.

(Refer Slide Time: 01:47)

Importance of the Assumptions

- These assumptions provide the theoretical basis for the t test and the F test used to determine whether the relationship between x and y is significant, and for the confidence and prediction interval estimates
- If the assumptions about the error term ζ appear questionable, the hypothesis tests about the significance of the regression relationship and the interval estimation results may not be valid.



5

Why this assumption is important these assumptions provide the theoretical basis for the t-test and F test used to determine whether the relationship between x and y is significant and for the confidence interval and prediction interval estimate. What we have done in the previous class we have done t test and F test to test these hypotheses what was our hypothesis: $\beta_1 = 0$ $\beta_1 \neq 0$. So, this assumption can be tested by two method only is the t-test and F test so to validate that assumptions the error about assumption is more important.

If the assumptions about the error term epsilon appear questionable the hypothesis test about the significance of the regression relationship and the interval estimation result may not be valid that is why we have to verify this assumptions.

(Refer Slide Time: 02:47)

Residuals for Ice cream parlours

Student Population x_i	Sales y_i	Estimated Sales $\hat{y}_i = 60 + 5x_i$	Residuals $y_i - \hat{y}_i$
2	58	70	-12
6	105	90	15
8	88	100	-12
8	118	100	18
12	117	120	-3
16	137	140	-3
20	157	160	-3
20	169	160	9
22	149	170	-21
26	202	190	12

Source: Statistics for Business & Economics, David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, Cengage Learning, 2013

6

We will take an example this example is adopted from statistics for Business and Economics David and Anderson Sweeney and Williams the student population x_i 2 6 8 8 12 and so on, sales of ice cream is given so we have fitted the regression line y_i \hat{y}_i equal to $60 + 5x_i$ so when you substitute the x value here this is actual 58 this is our predicted 70 the difference is $50 - 70$ is -12 , so 105, 90 the difference is 15, so this $y_i - \hat{y}_i$ is that is the residual. So this residual I have to have some properties that properties we will check it.

(Refer Slide Time: 03:41)

Residual analysis is based on an examination of graphical plots

- A plot of the residuals against values of the independent variable x
- A plot of residuals against the predicted values of the dependent variable \hat{y}
- A standardized residual plot
- A normal probability plot

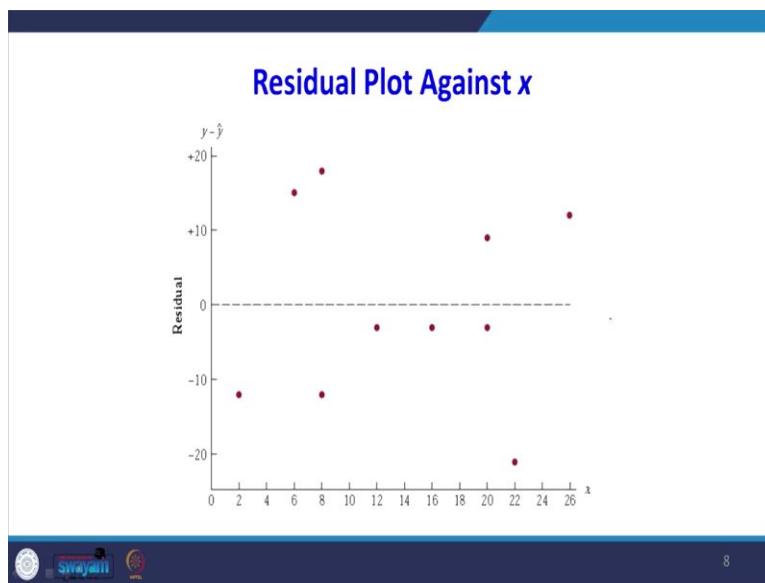
7

The residual analysis is based on the examination of graphical plot. So, we are going to plot the residual it was in the previous slide then we have to check certain assumptions there are 4 method we are going to do in this class one is a plot of the residual against value of independent

variable x. So, first assumption is x axis we are going to take x value in dependent variable in y axis we are going to take residual the second assumption is the plot of residuals against a predicted value of the dependent variable $y \hat{}$, so in x axis we are going to have $y \hat{}$ then y axis we are going to have residuals.

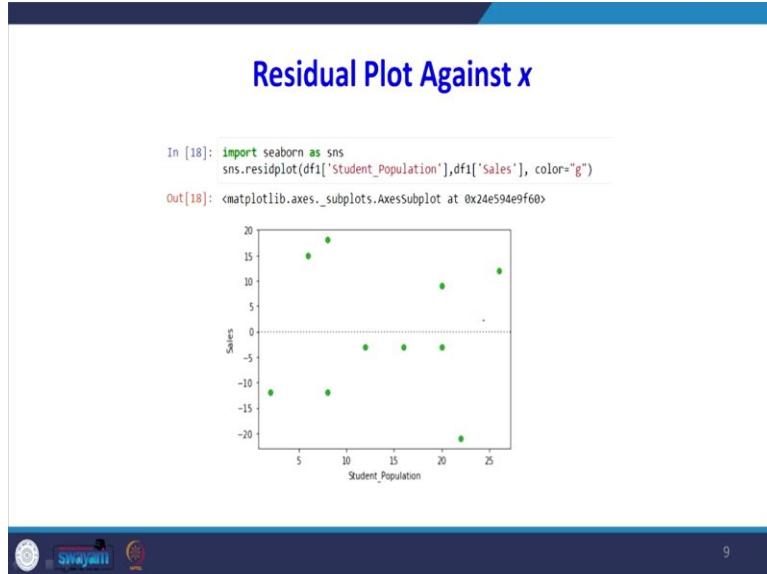
The third assumption is standardized as a residual plot we are going to standardize this the residual we know that how to standardize standardized for example if you are standardizing this is $(x - x \bar{}) / \sigma$ this is the way Sigma our standard error so it is nothing but z is nothing but $z = (x - x \bar{}) / \sigma$ so we will standardize our residual then we plot it the last assumption is normal probability plot.

(Refer Slide Time: 04:51)



So, we will check this assumptions the first assumption is when we plot the residuals against x value so this error that is duals are plotted in this way. So, what is the inference we can get it number one that it is not following any pattern, if it is not following any pattern these errors are independent then there is no problem in the assumptions.

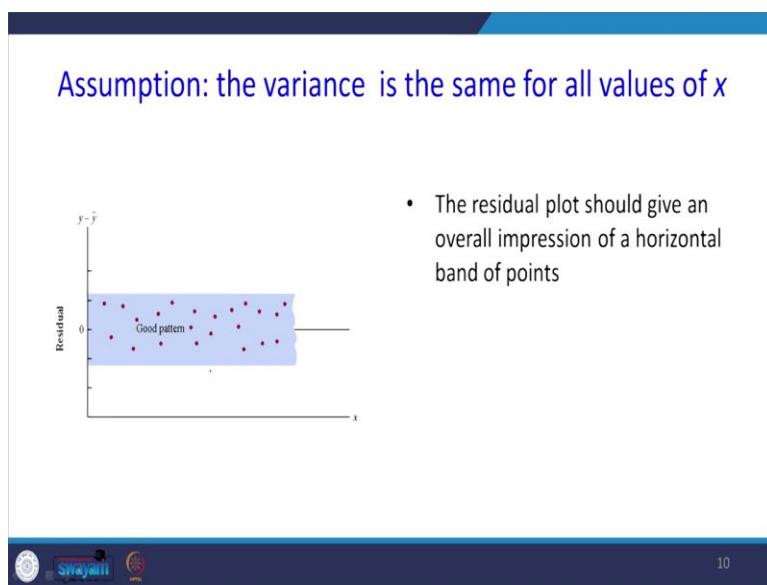
(Refer Slide Time: 05:15)



So, this one we have done with the help of Python I have the screenshot at the end of this lecture I am going to run all these codes you can verify it. Import seaborn as sns before that we have to import the data set that I will show you in the end of the class. So, sns.residplot so this is used for plotting that is do one variable a student population that is x value the y value sales color is green color so you will get this output.

So this was the Python output of a residual sales that is a y predicted value against sales y axis not the sales it is the residual, it is the residual, y axis is not the sales because sales would not become 0.

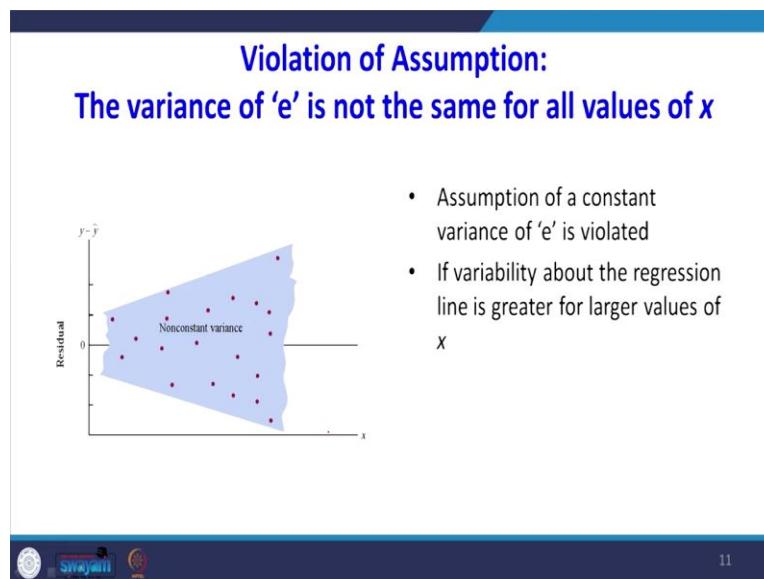
(Refer Slide Time: 06:20)



So, the one assumption is the variance is the same for all values of x , so what will happen now in this figure which is given it is looking like a rectangle shape that means this assumption is valid it is a good pattern. So, what did this graph implies the residual plot should give an overall impression of horizontal band of points. So, that means the variance even though the x value is increasing the variance is same so then we get here a horizontal band of points so this is the way to check the one assumption that the variance is same for all values of x .

Sometime what may happen when the value of x increases the variance may increase that should not be the case yeah that is an example of this one.

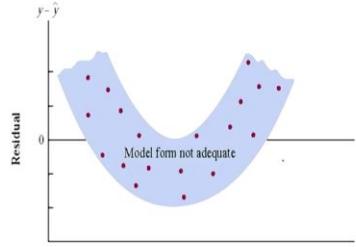
(Refer Slide Time: 06:54)



What is happening violation of assumption what is that the variance of y_i is not the same for all values of x when x is increases the variance is no it is getting a conical shape, so it is a non constant variance. This is the violation of our regression model. So, assumption of a constant variance of E is violated if you are getting this kind of shape. If variability about the regression line is greater for larger values of x then you can get this kind of pictures. So that is not correct one.

(Refer Slide Time: 07:30)

Assumed regression model is not an adequate representation

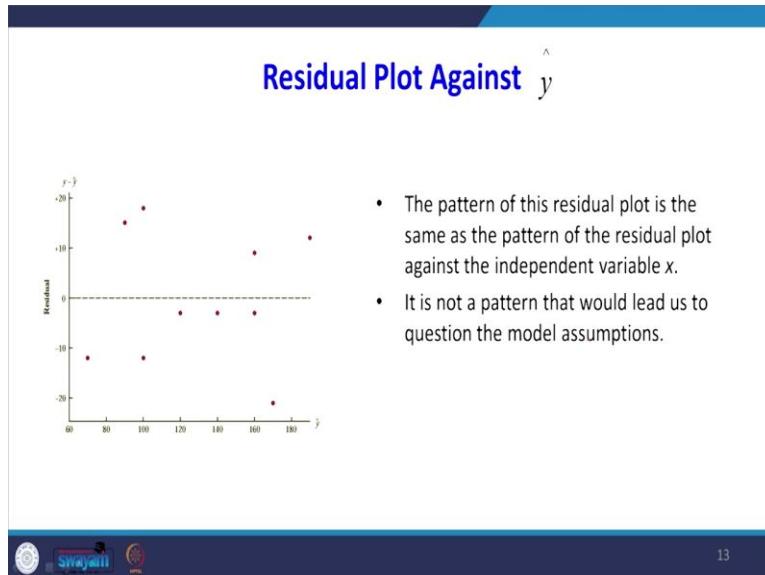


A curvilinear regression model or multiple regression model should be considered.

Another type of picture you may get it when you plot the residual against the x, a curvilinear this is a this is a kind of a non linear shape, so instead of fitting a linear regression equation it is suggesting that you can try for curvilinear regression model or a multiple regression model should be considered if you are getting the plot is in this shape. Previously we have plot the residual against x now we are going to plot the residual against \hat{y} that is our predicted value. The pattern of this residual plot is the same as the pattern of residual plot against an independent variable x.

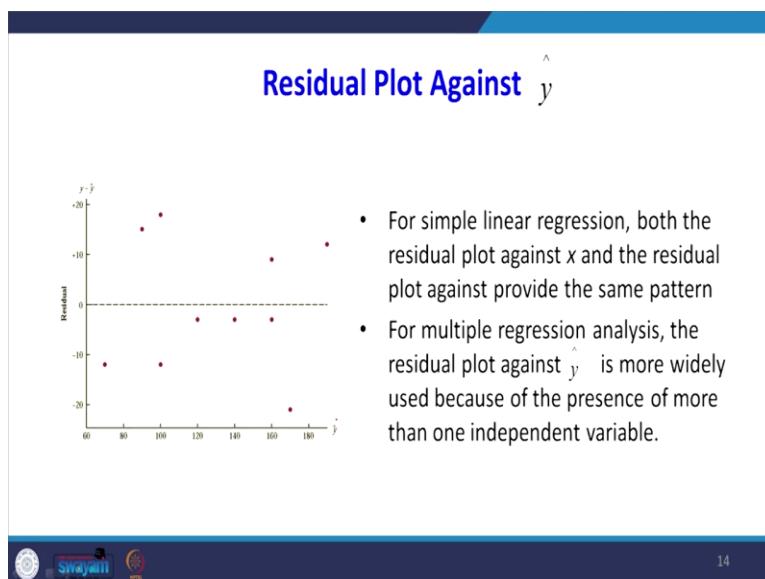
It is not a pattern that that would lead us to question the model assumptions why this we are going for that one if there are more number of independent variable for each independent variable you have plot this residual.

(Refer Slide Time: 08:19)



So, instead of going for different independent variable if you plot this residual against this predicted value then from that we can verify whether the model is valid or not.

(Refer Slide Time: 08:35)



So, for a simple linear regression both the residual plot against x and the residual plot against \hat{y} provided the same pattern for multiple regression analysis the residual plot against \hat{y} is the more widely used because of the presence of more than one independent variable. Whenever there is more than one independent variable instead of going for x we should go for \hat{y} .

(Refer Slide Time: 08:58)

Standardized Residuals

- Many of the residual plots provided by computer software packages use a standardized version of the residuals.
- A random variable is standardized by subtracting its mean and dividing the result by its standard deviation.
- With the least squares method, the mean of the residuals is zero.
- Thus, simply dividing each residual by its standard deviation provides the **standardized residual**



15

Then we will go for next residual plot that is a standardized residuals. Many of the residual plots provided by computer software packages uses a standardized version of the residuals. So, what is the standardized version yeah random variable is standardized by subtracting its mean dividing the result by its standard deviation, this way Z equal to a residual i value residual mean value divided by the standard deviation of the residual.

With the least square method the mean the residual is 0 because Sigma of $x - \bar{x}$ is 0, thus simply dividing each residual by its standard deviation provides the standardized to residual. So, what you have to do in the least square method simply how to divide residual by your standard deviation that will give you the standardized the residual.

(Refer Slide Time: 09:57)

Python Code

```
In [14]: import pandas as pd
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
import matplotlib.pyplot as plt

In [9]: df1 = pd.read_excel('icecream.xlsx')
df1

Out[9]:
   Student_Population  Sales
0                   2     58
1                   6    105
2                   8     88
3                   8    118
4                  12    117
5                  16    137
6                  20    157
7                  20    169
8                  22    149

In [11]: reg1 = ols(formula = "Sales ~ Student_Population", data = df1)
fitt1 = Reg1.fit()
print(fitt1.summary())

```

16

So, I am so in the Python code here this is a screenshot of the program import pandas as pd, from statsmodels dot formula dot api import OLS from stats model dot stat stat anova import anova underscore lm, import matplotlib dot pyplot as plt. So, the data set name is ice cream so in the independent variable student population the dependent variable is sales. So, to get your regression model reg1 equal to OLS formula equal to sales as a dependent variable tilde student underscore population data equal to df1.

(Refer Slide Time: 10:43)

OLS Regression Results

```
=====
Dep. Variable:      y      R-squared:   0.903
Model:              OLS      Adj. R-squared:  0.891
Method:             Least Squares      F-statistic: 74.25
Date: Thu, 05 Sep 2019      Prob (F-statistic): 2.55e-05
Time: 11:16:42      Log-likelihood: -39.342
No. Observations: 10      AIC: 82.68
Df Residuals: 8      BIC: 83.29
Df Model: 1
Covariance Type: nonrobust
=====
            coef  std err      t      P>|t|      [0.025      0.975]
-----
const    60.0000   9.226    6.503    0.000    38.725    81.275
x1       5.0000   0.580    8.617    0.000    3.662    6.338
=====
Omnibus: 0.928  Durbin-Watson: 3.224
Prob(Omnibus): 0.629  Jarque-Bera (JB): 0.616
Skew: -0.060  Prob(JB): 0.735
Kurtosis: 1.790  Cond. No. 33.6
=====
```

17

So, when you print a summary so you will get this kind of regression output. So, this says your r square this is over adjusted r square I will explain the meaning of our just r square in multiple

regression this was for F statistics. So, what say this is, so $y = 60 + 5x$ x is number of populations.

(Refer Slide Time: 11:07)

```
In [12]: print(anova_lm(Fit1))
```

	df	sum_sq	mean_sq	F	PR(>F)
Student_Population	1.0	14200.0	14200.00	74.248366	0.000025
Residual	8.0	1530.0	191.25	NaN	NaN

So, when you use this print anova_lm for the our model fit one you can get your ANOVA table for regression analysis, so for you a residual it is 8 because there is a 10 data set so the degrees of freedom is $n - p - 1$, p is number of independent variable there is only one independent variable so the degrees of freedom is 1 this is sum of square for student population this sum of square for error. So sum of squares divided by degrees of freedom you will get mean sum of square.

So the F value is nothing but means sum of squared divided by mean error sum of square so the p value is very low so we can say that the model is valid.

(Refer Slide Time: 11:52)

Standardized Residuals

STANDARD DEVIATION OF THE i th RESIDUAL

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i}$$

where

$s_{y_i - \hat{y}_i}$ = the standard deviation of residual i

s = the standard error of the estimate

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

19

Next will tell you how to find out these standardized to residual. So, the standardized residual is $s y_i - y \hat{i}$ equal to $s \sqrt{1 - h_i}$ here s is the standard error of the estimate. So, in the previous this is where MSE is 191 when you take the square root of this what is the standard error is standard error is SSE divided by $n - 2$, when you take square root otherwise 1 and 2 and 0.25 when you take square root that you will get the standard error.

So that is nothing but the value of s so you can find out h_i equal to $(1/n) + ((x_i - \bar{x})^2 / \sum(x_i - \bar{x})^2)$.

(Refer Slide Time: 12:41)

Computation of standardized residuals for Icecream parlors

Restaurant <i>i</i>	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$	h_i	$s_{y_i - \hat{y}_i}$	$y_i - \hat{y}_i$	Standardized Residual
1	2	-12	144	.2535	.3535	11.1193	-12	-1.0792
2	6	-8	64	.1127	.2127	12.2709	15	1.2224
3	8	-6	36	.0634	.1634	12.6493	-12	-.9487
4	8	-6	36	.0634	.1634	12.6493	18	1.4230
5	12	-2	4	.0070	.1070	13.0682	-3	-.2296
6	16	2	4	.0070	.1070	13.0682	-3	-.2296
7	20	6	36	.0634	.1634	12.6493	-3	-.2372
8	20	6	36	.0634	.1634	12.6493	9	.7115
9	22	8	64	.1127	.2127	12.2709	-21	-1.7114
10	26	12	144	.2535	.3535	11.1193	12	1.0792
	Total		568					

20

So, there is an illustration so I use there x is there we are finding $x_i - \bar{x}$ because this value will be useful for the the formula which is in the previous slide so $x_i - \bar{x}$ whole square so when you know we can $((x_i - \bar{x})^2 / \sum(x_i - \bar{x})^2)$, then you confront a h_i you can find out the $s_{y_i - \hat{y}_i}$ from that you can find out the $y_i - \hat{y}_i$, so then you will get the standardized residual.

(Refer Slide Time: 13:15)

Computation of standardized residuals for Icecream parlors

STANDARDIZED RESIDUAL FOR OBSERVATION i

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}}$$

21

It is do it so standardized residual is nothing but $y_i - \hat{y}_i / s_{y_i - \hat{y}_i}$ so what will happen when you plot this figure x against this and residual most of the data point is see that between + 2 and - 2 so that means that 95% of the time the data's are within the limit so this is acceptable.

(Refer Slide Time: 13:47)

Plot of The Standardized Residuals Against The Independent Variable x

```
In [21]: influence = fit1.get_influence()
resid_student = influence.resid_studentized_external

In [22]: resid_student
Out[22]: array([-1.09212653,  1.26798654, -0.94196706,  1.54023214, -0.21544891,
   -0.21544891, -0.22263461,  0.68766487, -2.01063738,  1.09212653])

In [24]: plt.figure()
plt.scatter(df1['Student_Population'], resid_student, color = "green")
Out[24]: cmatplotlib.collections.PathCollection at 0x24e5a382b38>
```

23

The assumption is valid we can plot the standardized residual plot against the independent variable x so for that you this command influence, influence equal to fit1 dot get underscore influence, resid_student equal to influence dot resid_studentized_external, external so you can see what is that resid_student that is an array of this is nothing but the standardized residual.

So now you can plot student population against this studentized residual will get this figure see all the data point is between + 2 and - 2, so this assumption is valid.

(Refer Slide Time: 14:22)

Studentized residual

- The standardized residual plot can provide insight about the assumption that the error 'e' term has a normal distribution.
- If this assumption is satisfied, the distribution of the standardized residuals should appear to come from a standard normal probability distribution.

24

The standardized residual plot can provide insight about the assumption that the error term 'e' has the normal distribution. If this assumption is satisfied the distribution of the standardized residual should appear to come from a standard normal probability distribution.

(Refer Slide Time: 14:41)

Studentized residual

- Thus, when looking at a standardized residual plot, we should expect to see approximately 95% of the standardized residuals between -2 and 2.
- We see in Figure that for the Ice-cream example all standardized residuals are between -2 and 2.
- Therefore, on the basis of the standardized residuals, this plot gives us no reason to question the assumption that 'e' has a normal distribution.



25

Studentized there is dual this when looking at the standardized the residual plot we should expect to see approximately 95% of the standardized residuals between - 2 and + 2. We see the figure that from the ice cream example all standardized residuals are between - 2 and + 2 therefore on the basis of the standardized residuals this plot gives us no reason to question that assumption that the error term has here normal distribution.

(Refer Slide Time: 15:10)

Normal Probability Plot

- Another approach for determining the validity of the assumption that the error term has a normal distribution is the **normal probability plot**.
- To show how a normal probability plot is developed, we introduce the concept of *normal scores*.



26

Next we will plot normal probability plot. Another approach for determining the validity of the assumption that the error term has a normal distribution is normal probability plot. Many software packages you may see that the normal probability plot. To show how a normal probability plot is developed we introduce concept called normal score. Suppose 10 values are

selected randomly from a normal probability distribution with the mean 0 and standard deviation 1 and that is sampling process repeated over and over with the values in each sample of 10 ordered from smallest to largest.

(Refer Slide Time: 15:52)

Normal Probability Plot

- Suppose 10 values are selected randomly from a normal probability distribution with a mean of zero and a standard deviation of one, and that the sampling process is repeated over and over with the values in each sample of 10 ordered from smallest to largest.
- For now, let us consider only the smallest value in each sample.
- The random variable representing the smallest value obtained in repeated sampling is called the first-order statistic.



27

For now let us consider only the smallest value in each sample. The random variable representing the smallest value obtained in a repeated sampling is called first order statistic. Okay the second largest is second order statistic and so on so this was the first order statistics.

(Refer Slide Time: 16:14)

Normal Probability Plot

NORMAL SCORES
FOR $n = 10$

Order Statistic	Normal Score
1	-1.55
2	-1.00
3	-.65
4	-.37
5	-.12
6	.12
7	.37
8	.65
9	1.00
10	1.55



28

So for this first order statistic wherein the sample size equal to 10 it should be - 1.55 that means these values data's are coming from the normal distribution. So, for the second order statistics the

value should be so, this value which we got from the table. Now we are going to compare the standardized residual values with this table so already we have the standardized residual when I equal to 1 x i equal to 2, so it is – 1.0792.

So these values we are going to compare with the standardized so this value we are going to compare with the normal scores.

(Refer Slide Time: 16:57)

Normal scores and ordered standardized residuals for Armand's pizza parlors	
Normal Scores	Ordered Standardized Residuals
-1.55	-1.7114
-1.00	-1.0792
-.65	-.9487
-.37	-.2372
-.12	-.2296
.12	-.2296
.37	.7115
.65	1.0792
1.00	1.2224
1.55	1.4230

So, what is happening here then we look at this picture when the order statistic is 1 the minimum value is – 1.55 so in this figure we have to see which is near to – 1.55, so the this one -1.71. So, this is the – 1.71 next we have to see which is the least value from this figure next to least value is 1.07 so – 1.07 next least. So, we have mapped with this our standardized the residual against the normal score. When it is the the least one is taken as – 1.71 when it is 1.55 in the normal score the corresponding score from our dataset is 1.4230.

(Refer Slide Time: 17:49)

Normal Probability Plot

- If the normality assumption is satisfied, the smallest standardized residual should be close to the smallest normal score, the next smallest standardized residual should be close to the next smallest normal score, and so on.
- If we were to develop a plot with the normal scores on the horizontal axis and the corresponding standardized residuals on the vertical axis, the plotted points should cluster closely around a 45-degree line passing through the origin if the standardized residuals are approximately normally distributed.
- Such a plot is referred to as a *normal probability plot*.



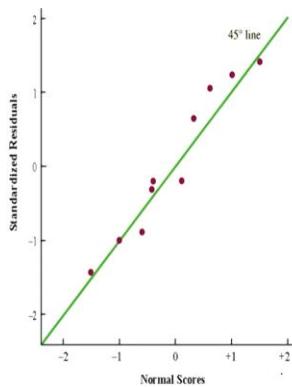
31

Now we will see what its normal probability plot by using this data we will plot it if the normality assumption is satisfied the smallest standardized the residual should be close to the smallest normal score. The next smallest standardized residual should be close to the next smallest normal score and so on that is what we mapped it in the previous slides.

If you had to develop a plot with the normal score on the horizontal axis and the corresponding standardized the residual on the vertical axis the plotted point should cluster closely around the affine is 45-degree passing through the origin. The standardized residuals are approximately normally distributed. It is a property of this residual plot. So, such a plot is referred as normal probability plot.

(Refer Slide Time: 18:36)

Normal probability plot for Ice Cream parlors

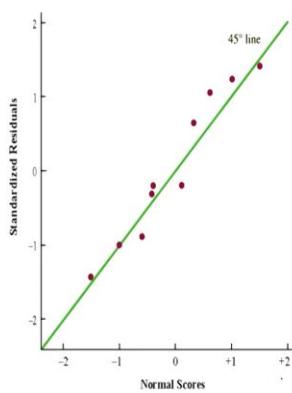


32

So, this was normal probability plot so in x-axis normal score is written in y-axis the standardized residuals written. So, it is starting from 0 the line is 45 degree so all the points are right it is not deviating from this line, so it is it is a clustering around this Green Line. So, then we can say that this data follow normal distribution. Suppose the data is following this by this point this point it is not going it is not clustered around this green line then we can say it is not following normal distribution.

(Refer Slide Time: 19:15)

Normal probability plot for Ice Cream parlors



32

This also we have done with the help of Python from `scipy import stat` `stats import stats` `import statsmodel` `dot api as sm` so we are going to take residual as `res` is equal to `fig 1 dot residual` then we go for `prop lot` equal to `sm dot probplot` because `sm` and all different library which is already important

to see that import statsmodels dot api as sm. So, then figure prompt lot dot qq plot when line equal to 45 degree, so we can say h= plt.title ('qq plot again to this dual of OLS. fit then we are getting this you see that all the points are above this red line then we can see this normality assumption is validated.

Now what we are going to do I have prepared this command in Python I go to run all the Python course then I am going to verify I am going to show how to get this residual plots then how to verify that it meet the regression assumptions. So, far I have shown the screenshot of the Python output now I am going to run and I am going to explain how to get the residual plots and what is the interpretation of that.

For that what I have done I have taken one regression example filenames where I have stored the data is ice cream so first I will run this I on the library then I will show what is the data set. This data set shows there is a two variable and this one a student underscore population is independent variable sales is dependent variable. So, for this data set I am going to run the regression equation. We are getting regression output you see that intercept is 60 the intercept of the student underscore population variable is 5.

So we can write it y equal to $60 + 5 \times 1$ here $x 1$ is student population then we can see that this p-value also it is less than 0.05 so this independent variable is significant values you see that r square is 0.903 that means 90.3% of the variability of Y is explained with the help of this is a regression model. Similarly for the x coefficient the standard error is 0.58. Now we are going to get the ANOVA table for this regression.

So this ANOVA table type this print anova underscore lm fit1 we are getting this anova table for regression analysis what we are understanding here for independent variable is student population, so the degrees of freedom is one sum of square is 14200 when you divide this 14200 by 1 we are getting the mean sum of square. Then for a error term the degrees of freedom is 8. How it is 8 because there was a 10 data set so the degrees of freedom is $10 - 1 - \text{number of independent variables}$ so $10 - 1, 9 - 1$ one independent variable so 8.

So the sum of square is 1530, so when you divide this 1532 by 8 we are getting the mean error sum of squares. So, F value is this 14200 by 191 you are getting this one so p value is very low this model is validated. Next what are you going to do we are going to draw the residual plot in x axis I have taken the student population that is independent variable in y axis this is not the sales it is the residual for the sales .

See, that next one we will see the studentized residual plot you run this, so this is your standardized residuals. So, we will plot this standardized residuals so this is a standardized visible so what is the interpretation from this is all the points are between + 2 and - 2 so we can say this assumption is valid. Next we will go for checking the normality of the error term. So now what is happening we are getting the qq plot when you run this code.

So this qq plot says that all the points are around this red line we can say this model is that is the assumption of the normality is tested it is correct. In every lectures you can follow this code you can verify this output I will also planning to share this code with you when you register this course. Now I will conclude what we have seen in this lecture in this lecture we have tested various assumptions about the regression models these assumptions we have tested with the help of different residual plots.

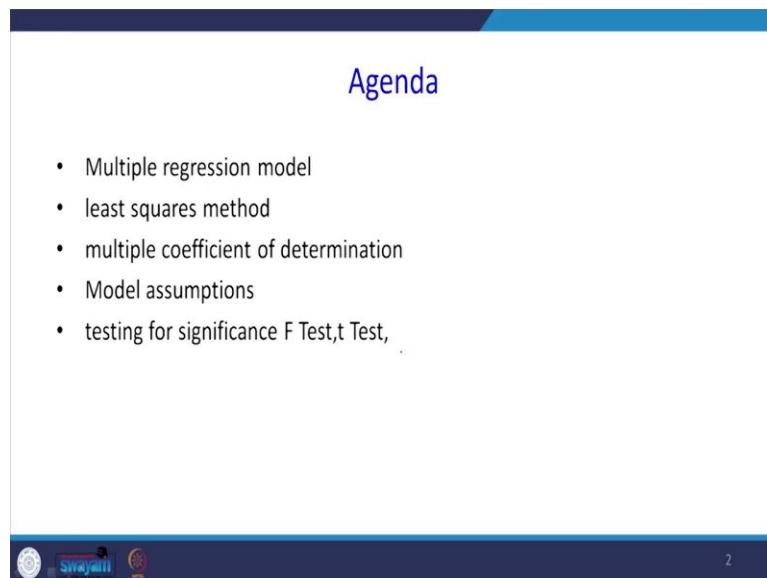
We have seen 4 types of residual plot 1 plot is a residuals against independent variable, the next one is the residual against our predicted values the third one is standardized the residual plot the fourth one is the normal probability plot. So, these different graphs helped us to test the assumption about the regression models. The next class we will discuss about the multiple regression models with some other examples, thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 33
Multiple Regression Model - I

In the previous class we have studied about simple linear regression, in this class we are going to discuss about multiple regression models.

(Refer Slide Time: 00:35)



The slide has a dark blue header and footer. The header contains the text 'Agenda'. The footer contains three small circular icons and the number '2'.

Agenda

- Multiple regression model
- least squares method
- multiple coefficient of determination
- Model assumptions
- testing for significance F Test,t Test,

The class agenda is I am going to explain what is multiple regression model then what is a least square method then multiple coefficient of determination. In the multiple coefficient of determination I am going to explain what is adjusted r-square also. Then what are the assumption in the multiple linear regression. Then I am going to test the significance of, by using F test and t test.

(Refer Slide Time: 01:04)

Multiple regression model

MULTIPLE REGRESSION MODEL

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

MULTIPLE REGRESSION EQUATION

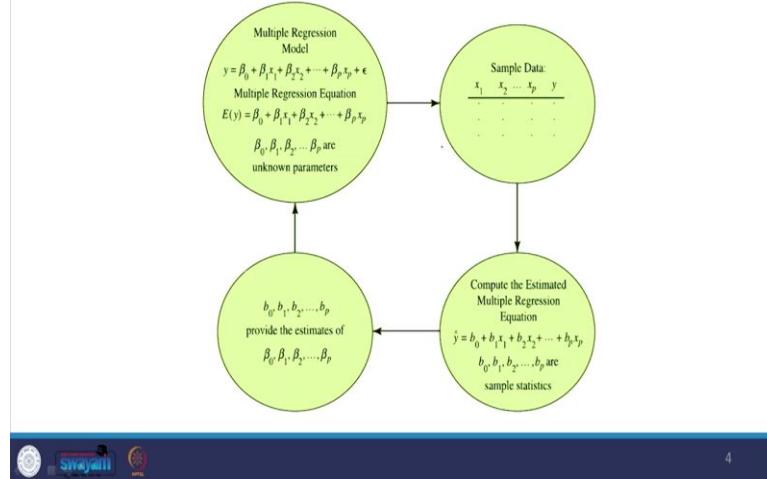
$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$



What is a multiple regression model so multiple regression model is when there are more than one independent variable that is called multiple linear regression model. If it is only one independent variable it is linear regression model. When you take the expected value of this multiple regression model so we know that that assumption any regression equations that the expected value of error term is 0, so when you take expected value of y there would not be any error term that is that is a multiple regression equation. Here β_1 β_2 is the coefficient of x_1 x_2 and β_p a coefficient of x_p .

(Refer Slide Time: 01:42)

The estimation process For multiple regression



What is the estimation process for a multiple regression there is a multiple regression model y equal to $\beta_0 \beta_1 \beta_2$ and β_p and e be an error term from this we can go for multiple

regression equations where β_0 β_1 β_2 are unknown parameters. To find out this unknown parameter from the population we are going to collect sample data for x_1 x_2 like this up to x_p and sample data for y that is dependent variable. With the help of sample data we are going to construct your sample regression equation what is that compute to the estimator multiple regression equation that is $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$ and so on $+ b_p x_p$ where b_0 b_1 b_2 b_p our sample statistics.

So with the help of sample statistics we are going to find out the population parameter that is β_0 β_1 β_2 β_p then there will do a significant test then we will see that whether the β_1 β_2 is equal to 0 or not equal to 0 after testing that we will find out what is the actual value of β_1 β_2 at the population level. This is the process of doing a multiple regression model. This is similar to the simple linear regression model but what we have done in the simple linear regression model only x_1 and y_1 was taken only one independent variable is there but here more than one independent variable that is only difference all other concepts are same.

(Refer Slide Time: 03:21)

Simple vs multiple regression

- In simple linear regression, b_0 and b_1 were the sample statistics used to estimate the parameters β_0 and β_1 .
- Multiple regression parallels this statistical inference process, with $b_0, b_1, b_2, \dots, b_p$ denoting the sample statistics used to estimate the parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_p$.

5

So, what is simple versus multiple regressions. In simple linear regression b_0 b_1 bear the sample statistics used to estimate the parameter of β_0 and β_1 but in multiple regression the parallel is that the statistical inference process with b_0 b_1 b_2 and b_p denoting the sample statics are used to estimate the parameter of β_0 β_1 β_2 and β_b . So, what is the meaning of this one is

with the help of sample statistics b 0 b 1 b 2 we are going to predict the population parameter the beta 0 beta 1 and beta 2.

In simple regression there was only b 0 was there beta 1 was there only one independent variable in multiple regression more than one independent variable there is only difference.

(Refer Slide Time: 04:18)

Least Squares Method

LEAST SQUARES CRITERION

$$\min \sum (y_i - \hat{y}_i)^2$$

6

Least square method in simple linear regression also I have derived the formula for b 0 b 1 by having the assumption that when we draw a line the error term that is the sum of the square of the error has to be minimized. But the \hat{y} there in simple linear regression \hat{y}_i was $b_0 + b_1 x_1$ but in multiple regression this \hat{y}_i equal to $b_0 + b_1 x_1 + b_2 x_2$ and so on + $b_p x_p$, p is the number of independent variable.

So, all other procedure is same here also what we are going to do that there are but here it is a multi-dimensional picture we cannot draw a two-dimensional picture because we need it because there are more than one independent variable that is going to be a a multi-dimensional picture that we cannot explain with the help of a simple graph.

(Refer Slide Time: 05:17)

Least Squares Method

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

7

The least square estimate what happened to \hat{y} equal to $b_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ and x_p because there would not be error term here because the expected value of the error term becomes 0. So, how to interpret the value of β_1 , β_2 and β_3 how do you interpret the coefficient of β_1 is by keeping other variables constant if the x_1 is improved by one unit the \hat{y} will be improved by β_1 units. It is a similar way for simple linear regression but here when you are interpreting one coefficient we have to assume that the other coefficient for other independent variables are constant.

(Refer Slide Time: 06:03)

An Example: Trucking Company

- As an illustration of multiple regression analysis, we will consider a problem faced by the Trucking Company.
- A major portion of business involves deliveries throughout its local area.
- To develop better work schedules, the managers want to estimate the total daily travel time for their drivers.

Source: Statistics for Business and Economics, 2012, Anderson

8

We will take an example this example problem is taken from statistics for Business and Economics is the author by Andersen. As an illustration of multiple regression analysis we will

consider a problem faced by a tracking company the major portion of the business involves deliveries throughout the local area to develop a better work schedule the manager want to estimate total daily travel time for their drivers. So, they want to estimate this is going to be total daily travel time is going to be our dependent variable.

(Refer Slide Time: 06:43)

PRELIMINARY DATA FOR BUTLER TRUCKING		
Driving Assignment	$x_1 = \text{Miles Traveled}$	$y = \text{Travel Time (hours)}$
1	100	9.3
2	50	4.8
3	100	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1

There are 10 assignments there are 10 assignment drivers x 1 equal to miles traveled y equal to travel time there is a connection between x 1 and y what is the meaning of that 1 when the travel time we will increase distance traveled also high. So, y is the dependent variable x 1 is independent variable.

(Refer Slide Time: 07:03)

Using python import data

```
In [1]: import pandas as pd
         from statsmodels.formula.api import ols
         from statsmodels.stats.anova import anova_lm
         import matplotlib.pyplot as plt

In [2]: df1 = pd.read_excel('Trucking.xlsx')
        df1
```

I have brought the screenshot at end of this lecture I will run this quotes then you can understand it better when I will show that I will explain the screenshot import pandas as pd from statsmodels dot formula dot api import Wireless that is ordinary least square regression models from stats model dot stats dot anova import anova underscore lm because this library will be used to see the ANOVA table for a regression model.

Then import matplotlib dot pyplot as a plt the file name is which I have stored is it tracking that is an excel file I going to store this data into an object called df1, df1 equal to pd dot read underscore excel that file name, so if you want to know what is the data set this is the data set.

(Refer Slide Time: 07:54)

The screenshot shows a Jupyter Notebook cell with the title "Using python import data". The code cell contains the following Python code:

```
Out[2]:
```

	Driving Assignment	x1	n_of_deliveries	travel_time
0	1 100		4	9.3
1	2 50		3	4.8
2	3 100		4	8.9
3	4 100		2	6.5
4	5 50		2	4.2
5	6 80		2	6.2
6	7 75		3	7.4
7	8 65		4	6.0
8	9 90		3	7.6
9	10 90		2	6.1

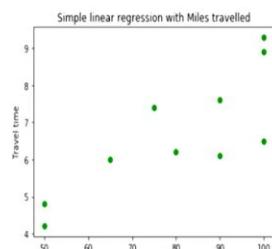
The notebook interface includes a toolbar with icons for file operations, a search bar, and a cell identifier "11" in the bottom right corner.

So, in this data set there are 1 travel underscore time is dependent variable there are 2, independent variable one is x 1 and another is number of deliveries. The meaning of x 1 is miles traveled before going to regression first we ought to have an idea between this independent variable x 1 miles traveled and time dependent variable is there any connection.

(Refer Slide Time: 08:19)

Scatter Diagram Of Preliminary Data For Trucking x_1

```
In [3]: import matplotlib.pyplot as plt  
plt.scatter(df1['x1'],df1['travel_time'], color = "green")  
plt.ylabel('Travel time')  
plt.title(' Simple linear regression with Miles travelled ')
```



12

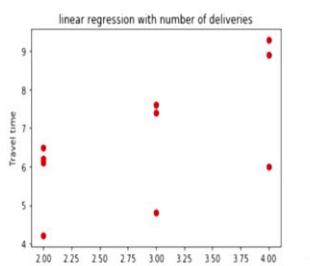
So the first step is first you have to draw the scatter plot so import the matplotlib dot pyplot as a plt, I am drawing the scatter plot df1 x 1 is in the x-axis travel underscore time in y-axis green color so label is travel time this one, so what is happening that there seems to be some relation between this miles traveled and the travel time that means the obviously when the miles traveled is more the travel time also will be more. This is between one independent variable and one dependent variable.

(Refer Slide Time: 08:55)

Scatter Diagram Of Preliminary Data For Trucking x_2

```
In [11]: plt.scatter(df1['n_of_deliveries'], df1['travel_time'], color = "red")  
plt.ylabel('Travel time')  
plt.title('linear regression with number of deliveries')
```

Out[11]: Text(0.5,1,'linear regression with number of deliveries')

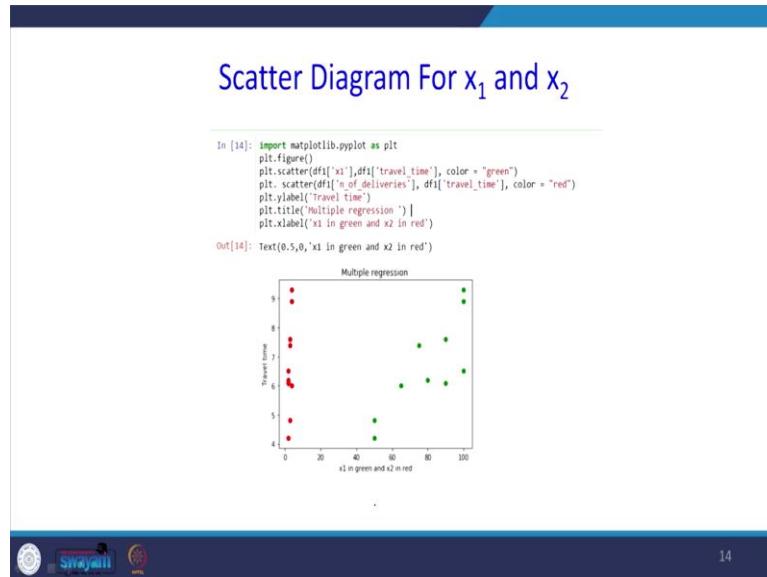


13

Similarly we will take another variable number of deliveries as an independent variable then travel time as the dependent variable there also seems to be there is a positive correlation. Why it

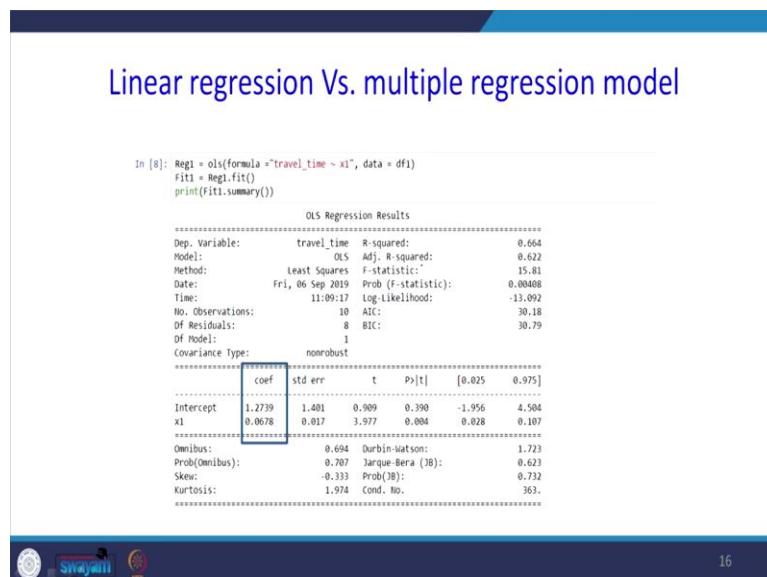
is required that if there is no correlation at all between that independent variable and dependent variable we need not do the regression analysis.

(Refer Slide Time: 09:20)



Now in this graph both the variables are taken together what is that vary the distance traveled and the number of deliveries this is the code for to show both variables in the same figure. So, what are I'm going to do first I am going to take one independent variable I am going to plot, construct the regression equation then I am going to take both intermediate variables together then I go to construct a regression equation. The first taking for one independent variable this is a $y \hat{}$ equal to $1.27 + 0.0678 x_1$, I will show you in the next slide how we got this answer.

(Refer Slide Time: 10:01)



So I am going to do a regression analysis that regression model I am going to say reg1 is equal to OLS formula the travel time is taken as the dependent variable x 1 distance traveled is taken as the independent variable. So, fit1 equal to reg1 dot fit so print fit1 dot summary so what is happening here we are getting the coefficient what is it coefficient the intercept is 1.2739 x1 is 0.0678, so how we can write it $y \hat{}$ equal to $1.2739 + 0.0678 x_1$ variable this is an independent variable with you see that the same answer we are getting here.

So for here one more things I were to understand see the R square is 0.664 okay now the next one what I am going to do I am going to introduce another variable here after introducing the another variable I am going to see what is going to happen this r square. The r square says the goodness of the model the higher the r square the model is better what is the meaning of 66.4 here was 0.664, 66.4% of the variability of y can be explained with the help of this model.

(Refer Slide Time: 11:25)

Linear regression Vs. Multiple regression model

- Multiple regression

$$\hat{y} = - .869 + .0611x_1 + .923x_2$$

Now what happening that I am going to bring another independent variable that is number of deliveries so when you bring another independent variable I will show you that model you see that model equal to OLS(travel underscore time tilde sign x 1 + n underscore of underscore deliveries so this is two independent variable if there are three you can write it plus that variables this is the way to do the multiple regression in Python.

(Refer Slide Time: 11:46)

Linear regression Vs. Multiple regression model

```
In [15]: from statsmodels.formula.api import ols
model = ols('travel_time ~ n_m_of_deliveries', data=df).fit()
model.summary()

C:\Users\vp\Anaconda\lib\site-packages\scipy\stats\stats.py:1390: UserWarning: "anyarray, n=21" % int(n))
  OLS Regression Results
==============================================================================
Dep. Variable: travel_time   R-squared:      0.904
Model: OLS   Adj. R-squared:   0.876
Method: Least Squares   F-statistic:    32.88
Date: Fri, 06 Sep 2019   Prob (F-statistic):  0.000276
Time: 11:10:53   Log-Likelihood: -6.898
No. Observations: 10   AIC: 19.68
Df Residuals: 7   BIC: 20.59
Df Model: 2
Covariance Type: nonrobust
==============================================================================
            coef  std err      t  P>|t|  [0.025  0.975]
Intercept  0.8687  0.952  0.913  0.392  3.119  1.381
x1         0.0611  0.019  3.162  0.000  0.038  0.085
n_m_of_deliveries  0.9234  0.221  4.176  0.004  0.401  1.446
Omnibus: 0.039  Durbin-Watson: 2.515
Prob(Omnibus): 0.981  Jarque-Bera (JB): 0.151
Skew: 0.074  Prob(JB): 0.927
Kurtosis: 2.418  Cond. No. 435

```



18

So, now what is happening here you look at the y intercept it is $y = -0.8687 + 0.0611 \times 1 + 0.9234 \times 2$, here you can call it as x_2 is what is the meaning of x_2 number of deliveries okay so, what is this, this is important. We will verify this in the previous slide also we got the same thing $-0.869 + 0.0611 \times 1 + 0.923 \times 2$ now look at this the previous r square now look at this now this r square after introducing new variable.

After introduce a new variable the r square is previously it was 0.6 something now it is increased to 0.90 so adding a new variable as helped to improve the explaining power of this regression model. Then I explain there is one more term adjusted r-square because in many previous lectures I am saying that I will do the next lecture but I am not able to do that one now in this lecture I will explain what is the meaning of adjusted r-square.

The other point you have to understand you look at the p-value for each independent variable. So, what is the null hypothesis for a year what is the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$, so in all hypothesis for you look at the b values here see for x_1 it is a point 0 0 so we have to reject null hypothesis. When you reject null hypothesis $\beta_1 \neq 0$, that mean there is a relation between x_1 and y_1 .

Similarly look at the number of deliveries corresponding p-value is 0.04 that also less than 0.05 so that hypothesis $\beta_2 = 0$ also at we rejected that means at a population level there is a the

relationship is significant what is the meaning of that one is even at the population level between x 2 and y there is a significant relationship is there.

(Refer Slide Time: 14:04)

Multiple Coefficient of Determination

RELATIONSHIP AMONG SST, SSR, AND SSE

$$SST = SSR + SSE$$

where

$$SST = \text{total sum of squares} = \sum(y_i - \bar{y})^2$$
$$SSR = \text{sum of squares due to regression} = \sum(\hat{y}_i - \bar{y})^2$$
$$SSE = \text{sum of squares due to error} = \sum(y_i - \hat{y}_i)^2$$

19

Relation among SST SSR and SSE we know that SST total sum of square equal to the regression sum of square + error sum of square, SST this I have explained in my previous lecture total sum of square is this way for your convenience I am drawing one more time this is your \bar{y} bar this is your \hat{y} so this is y , so this distance okay this distance is your SSR this distance is your SSE, so the total distance is SST.

So this total distance is SST, so, what is SST? SST is $y_i - \bar{y}$ whole square Sigma what is SSR $\hat{y} - \bar{y}$ whole square what is SSE $y - \hat{y}$ whole square so when we have only one independent variable look at this here what is SST when you add this SST equal to summation of 15.87 + 8.02 so it will come around 23.89 SST. You see the residual sum of square so what is SSE? SSE is 8.02 when there is only one independent variable SSR is 15.871 to get this regression model output you have to use this one print anova _lm the or to call the first regression model.

(Refer Slide Time: 16:08)

Multiple Coefficient of Determination for Multiple regression model

```
In [18]: anova_table = anova_lm(model, typ=1)
anova_table
```

Out[18]:

	df	sum_sq	mean_sq	F	PR(>F)
x1	1.0	15.871304	15.871304	48.315660	0.000221
n_of_deliveries	1.0	5.729252	5.729252	17.441075	0.004157
Residual	7.0	2.299443	0.328492	NaN	NaN



21

The next slides we are going to bring another ANOVA table when there are two independent variables for that purpose and I want a score table equal to `anova_lm(model, typ=1)` ANOVA table. Now you see that the SST is same SST is around 22 around 22 but look at SSE is 2.29 so error has been decreased. You see SSR, SSR is these two 15.87 + 5 approximately 20. Something so what is happening when you introduce a new variable the value of SSR is increased to 20 variously SSR only one independent variable SSR is 15.

So after introduced a new variable the 5 unit of variants is increased and at the other point is previously when there are only one independent variable is the error term is 8.02. Now the error is reduced to 2.29 so that is the advantage of using more number of independent variable so that we can have more accurate model.

(Refer Slide Time: 17:20)

Multiple Coefficient of Determination

MULTIPLE COEFFICIENT OF DETERMINATION

$$R^2 = \frac{SSR}{SST}$$

$$R^2 = \frac{21.601}{23.900} = .904$$



22

Now you will see what is multiple coefficient of determination when there is a simple linear regression model we have called it coefficient of determination. Now there is a multiple independent variable we are going to call it is multiple coefficient of determination it is SSR by SST. So, what is R square SSR, SSR is when you add this two $15.87 + 5.7$, 21.6 . SST is when you are all three 22.2 approximately 23.0 .

So there is a 90.4% of the variability of y can be explained with the help of these two independent variable. So, the r square is increased so it is a good model when compared to simple linear regression model.

(Refer Slide Time: 18:07)

Multiple Coefficient of Determination

- Adding independent variables causes the prediction errors to become smaller, thus reducing the sum of squares due to error, SSE.
- Because $SSR = SST - SSE$, when SSE becomes smaller, SSR becomes larger, causing $R^2 = SSR/SST$ to increase.
- Many analysts prefer adjusting R^2 for the number of independent variables to avoid overestimating the impact of adding an independent variable on the amount of variability explained by the estimated regression equation.



23

So, now we will go for another concept adjusted R square what is the purpose of adjusted R square. So, adding independent variable causes the prediction errors to become smaller, so we know that see SST equal to SSR + SSE so when you add independent variable prediction error become smaller what will happen this error will become smaller so what will happen this when SSE becomes smaller SSR will become bigger one because SSR equal to SST - SSE when SSE becomes smaller SSR become larger.

So, causing R square to increase whenever you add any independent variable SSR will increase SSE will decrease due to that SSR will increase due to that the R square will increase. Many analysts prefer adjusting R square for number of independent variable to avoid overestimating the impact of adding an independent variable on the amount of variability explained by the estimated regression equation.

So what is happening instead of using R square we are going for adjusted R square. The advantage of adjusted R Square is whether the added new variable is it is really as an explaining variable or it is a noise variable otherwise the added a new variable how much it is helping to explain the variance of the existing model.

(Refer Slide Time: 19:38)

Adjusted Multiple Coefficient of Determination

n = number of observations
 p = denoting the number of independent variables

ADJUSTED MULTIPLE COEFFICIENT OF DETERMINATION

$$R_a^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

$$R_a^2 = 1 - (1 - .904) \frac{10 - 1}{10 - 2 - 1} = .88$$


SWAYAM
WORKSHOP

24

So, what is a formula for adjusted R square s previously what was the formula for R square see that R square equal to SSR divided by SST, explained variance divided by overall variance. So,

this explained variance the regression sum of square can be written this way $SST - SSE$, $SST - SSE$ because what is happening this SSR this SSR represents regression sum of square for all independent variables. So, when you add a new variable you cannot know the contribution of that new variable into the SSR we are going to split this SSR into two term that is $SST - SSE$ so now this will become $1 - SSE$ divided by SST .

But what we have to do we have to write the degrees of freedom because what is the meaning of adjusted is this adjusting for degrees of freedom. so, when SSE what is the degrees of freedom SSE the degrees of freedom is $n - p - 1$ what is the n , n is the total number of data set p is number of independent variable - 1 here it will become $n - 2$ divided by SST you write SST as it is. It is $n - 1$, so when you simplify this you will get this method.

So here what is the n , n is number of observations what is the p it is number of independent variables and you substitute here R^2 equal to $1 - (1 - R^2) / ((n - 1) / (n - p - 1))$ when you expand this R^2 otherwise you write R^2 equal to SSR / SST you will end up with this relationship this is adjusted R^2 is 0.88.

(Refer Slide Time: 21:36)

```
In [15]: from statsmodels.formula.api import ols
model = ols('travel_time ~ n_of_deliveries', data=df1).fit()
model.summary()

C:\Users\HP\Anaconda3\lib\site-packages\scipy\stats\stats.py:1390: UserWarning
  &amp;anyway, n>10
  &amp;anyway, n>1 % int(n))

Out[15]: OLS Regression Results
Dep. Variable: travel_time   R-squared:  0.884
Model: OLS                   Adj. R-squared:  0.876
Method: Least Squares        F-statistic: 32.88
Date: Fri, 01 Sep 2017   Prob(F-statistic): 0.000276
Time: 11:18:53   Log-Likelihood: -6.8398
No. Observations: 10   AIC: 19.68
Df Residuals: 7   BIC: 20.59
Df Model: 2
Covariance Type: nonrobust

coef std err      t  P>|t|  [0.025  0.975]
Intercept  0.8867  0.052  16.93  0.000  -3.119  1.881
x1         0.0011  0.010  0.102  0.000  -0.038  0.046
n_of_deliveries  0.0234  0.021  4.176  0.004  0.401  1.446

Omnibus: 0.039  Durbin-Watson: 2.555
Prob(Omnibus): 0.981  Jarque-Bera (JB): 0.151
Skew: 0.074  Prob(JB): 0.927
Kurtosis: 2.418  Cond. No. 435
```

You look at this that is the meaning of 0.8, so another importance of this adjusted R^2 is sometime you see what will happen I am writing here R^2 adjusted R^2 what will happen whenever you introduce new variable the value of R^2 will increase adjusted R^2

square also will increase. So, I will explain what is the meaning of R square and I just R square assume that there is a one dependent variable there are many independent variable that independent variable is x 1 x 2 x 3 and x 4.

Now what I am doing here I am going to build a regression model. So, first what I will do first I will take y then I will write regression equation in terms of x1 so what will happen R Square increase and will also adjusted R square. Now taking y is a dependent variable I am going to bring 2 independent variable R square increases adjusted R square also will increase. So, what will happen if the x2 is really helping to explain the variance of the y some time suppose say variable x 3, x 1, x 2 this x 3 variable is the noise variable.

Noise variable means it will not help to explain the variability why it is going to disturb the existing relationship. So, what will happen R square will increase adjusted R square will start decreasing. So, this is the hint for us that the variable which you have added is not helping to explain the model instead of that it is deteriorating the existing model. So, x 3 should not be added that is the meaning of this adjusted R square most of the time.

If the value of R square adjusted R square is similar that means that we have no need to increase any further variable into the model that means you have reached the good model.

(Refer Slide Time: 23:30)

Adjusted Multiple Coefficient Vs Multiple Coefficient

- If a variable is added to the model, R^2 becomes larger even if the variable added is not statistically significant.
- The adjusted multiple coefficient of determination compensates for the number of independent variables in the model.

If there is a gap for example R square is 0.9 adjusted R square is 0.3 that there is a possibility of adding more independent variable into the model. Now let us see adjust and multiple coefficient was this multiple coefficient if your variable is added to the model yeah that is the point which I am saying previous slide, if your variable is added to the model R square become larger even if the variable added is not statistically significant it is very important.

The adjusted multiple coefficient of determination that is adjusted R square compensate for the number of independent variable. So, it is adjusted means it is adjusted for the number of independent variable otherwise adjusted for it is a degrees of freedom. If the value of R square is smaller and the model contains a large number of independent variable adjusted total coefficient of determination can take negative value.

It is a very important point here the interpretation of R square and adjusted R square is not same. The R square is that how much variability of y is explained but the adjusted R square is not the same interpretation. What will happen many time adjusted R square may become negative okay you should be very careful on that. Then we will go for checking model assumptions so as I told you in the beginning of the class y equal to this is the regression model when you there will be error term when you go for regression equation there would not be error term because when you go for expected value of y $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ and so on.

And there would not be error term because the expected value of error is 0. We will go for some assumption what is the first assumption the error term epsilon is a random variable with mean or expected value of 0 what is implication for the given value of $x_1 x_2$ and up to x_p the expected or average value of y is given by this way you look at this when you go for expected value of y there is no error term. This equation represents the average of all possible values of y that might occur for the given value of $x_1 x_2$ up to x_p by expected value of y.

(Refer Slide Time: 25:58)

Assumption about error term

2. The variance of ϵ is denoted by σ^2 and is the same for all values of the independent variables x_1, x_2, \dots, x_p .

Implication: The variance of y about the regression line equals σ^2 and is the same for all values of x_1, x_2, \dots, x_p .

3. The values of ϵ are independent.

Implication: The value of ϵ for a particular set of values for the independent variables is not related to the value of ϵ for any other set of values.

4. The error term ϵ is a normally distributed random variable reflecting the deviation between the y value and the expected value of y given by $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$.

Implication: Because $\beta_0, \beta_1, \dots, \beta_p$ are constants for the given values of x_1, x_2, \dots, x_p , the dependent variable y is also a normally distributed random variable.



30

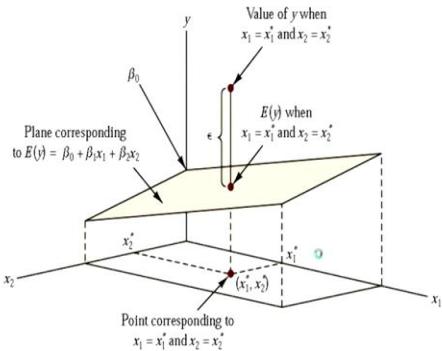
We will go for second assumption the variance of epsilon is denoted by Sigma square and is the same for all values of the independent variable $x_1 x_2 x_p$ what is implications the variance of y about the regression line equal to Sigma square and is the same for all values of $x_1 x_2 x_p$. if it is different we will call it is there is effect of heteroscedasticity. Why this point is required if you want to compare the variance of $x_1 x_2$ up to x_p should be same then only there is a meaning for comparison.

The third assumption is the value of epsilon are independent. What is implications the value of epsilon for a particular set of values for independent variable is not related to the value of epsilon for any other set of values. Another way the error terms are independent when you plot that error term there should not be any pattern whether it is increasing or decreasing pattern that is the meaning of this third assumption. Then fourth assumption the error term epsilon is normally distributed random variable reflecting the deviation between y value and the expected value of y given by $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ up to $\beta_p x_p$.

What is implications because of $\beta_0 \beta_1 \beta_p$ are constant for given values of $x_1 x_2 x_b$ the dependent variable y also normally distributed random variable because what will happen the error term it should be independent but it should follow a normal distribution with equal variance if it is not equal variance then it will go to the second assumptions also get violated.

(Refer Slide Time: 27:39)

Graph of the regression equation for multiple regression analysis with two independent variables



31

Now look at this graph of a regression equation for multiple regression analysis with 2, independent variable x_1 is one independent variable x_2 is another independent variable. See this is the mean value of x_1 this is mean value of x_2 you see this is a plane. So, multiple regression equation is explained with the help of here a surface otherwise this is called a surface the reference model is a plane now the equation is not the line it is the plane.

Otherwise they will call it is RSM also response surface model another name for regression is response surface model because now this is the surface.

(Refer Slide Time: 28:26)

Response variable and response surface

- In regression analysis, the term response variable is often used in place of the term dependent variable.
- Furthermore, since the multiple regression equation generates a plane or surface, its graph is called a response surface.

32

A response variable and response surface in regression analysis the term response variable is often used in place of the term dependent variable instead of saying dependent variable we will say the response variable. Furthermore since the multiple regression equation generates a plane or surface the graph is called response surface. In this lecture I have explained what is a multiple regression model? Then I have explained what is the connection between simple linear regression model and multiple regression model.

Then I explained the least square model then I have explained what is the meaning of R square and adjusted R square? Then I have explained various model assumptions. The next lecture I am going to test the significance of beta 1 beta 2 and beta 3 with the help of F test and t test and also we will see a demo on Python programming to do a multiple regression, thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 34
Multiple Regression Model - II

In the previous lecture we started multiple regression models. In the multiple regression model I have explain how to do a multiple regression model. What is the meaning of beta 0 beta 1 beta 2 and also explain what R squared and adjusted R square. In this lecture you are going to see how to do the significance test that means here also, like simple regression we are going to have some hypothesis about beta 1 coefficient and beta 2 coefficients so on.

And we are going to test whether the beta 1 is equal to 0 are not equal to 0. So what you are going to do in this lecture is we are going to test the significance of regression model with the help of F test and t test and I am going to do a Python demo for a multiple regression.

(Refer Slide Time: 01:13)

Testing for Significance

- The F test is used to determine whether a significant relationship exists between the dependent variable and the set of all the independent variables; we will refer to the F test as the test for overall significance.
- If the F test shows an overall significance, the t test is used to determine whether each of the individual independent variable is significant.
- A separate t test is conducted for each of the independent variables in the model; we refer to each of these t tests as a test for individual significance.

3

F test is used to determine whether a significant relationship exists between the dependent variable and the set of independent variables we will refer F test is the test of overall significance. I will show you where this F test is appearing in our Python output. If the F test shows an overall significance the t test is used to determine whether each of the individual independent variable is significant or not. A separate t test is conducted for each of the dependent

variable in the model. So we refer each of these, t-test as a test of individual significance. So, F test is used for testing the overall model of regression equation that t test is used to test for individual independent variables, whether they are significant are not.

(Refer Slide Time: 02:08)

F Test

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

The hypotheses for the F test involve the parameters of the multiple regression model.

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a: \text{One or more of the parameters is not equal to zero}$$

4

The F test what is the null hypothesis, here null hypothesis beta 1 equal to beta 2 up to beta p equal to 0 when I accept null hypothesis, what is the meaning is for example for accept beta 1 equal to 0 there is no relation between x1 coefficient and the different variable y. If I accept to beta 2 equal to 0 there is no relation between x 2 and y variable obviously alternative hypothesis is one or more of the parameter is not equal to 0.

(Refer Slide Time: 02:43)

F test significance

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a: \text{One or more of the parameters is not equal to zero}$$

TEST STATISTIC

$$F = \frac{\text{MSR}}{\text{MSE}}$$

REJECTION RULE

<i>p</i> -value approach: Reject H_0 if <i>p</i> -value $\leq \alpha$ Critical value approach: Reject H_0 if $F \geq F_\alpha$

where F_α is based on an *F* distribution with p degrees of freedom in the numerator and $n - p - 1$ degrees of freedom in the denominator.

5

So how to find out the F statistic of statistics is MSR divided by MSE that is mean regression sum of square divided by mean error sum of square how we are getting this mean regression sum of square when you divide SSR divided by corresponding degrees of freedom here the degrees of freedom p divided by then will get MSR. MSE when you divide SSE divided by $n - p - 1$ where p is number of independent variable. Then we will get the mean error sum of square.

This hypothesis testing can be done by two way one is by p value approach and release by critical value approach. In the p value approach reject H_0 if the p value is less than or equal to alpha what will happen this is F test. This way it is a right skewed data. This is your alpha value will get F alpha corresponding this value can get it in table. What we have to do you have to find out the p-value. The p-value is lying on the rejection side you have to reject it.

Otherwise, if the F value is beyond the alpha value we have to reject it where F alpha is based on the F distribution with p degrees of freedom in the numerator and $n - p - 1$, degrees of freedom in the denominator.

(Refer Slide Time: 04:18)

F test significance

$$F = \frac{10.8}{.328} = 32.9$$

So we got F equal to 10.8 divided by 0.328 equal to 32.9 I will show in another table this F value the previous slide.

(Refer Slide Time: 04:34)

F Test

```
In [15]: from statsmodels.formula.api import ols
model = ols('travel_time ~ n_no_of_deliveries', data=df).fit()
model.summary()

C:\Users\HP\Anaconda3\lib\site-packages\scipy\stats\stats.py:1990: UserWarning: anyarray, n=1" % int(n))

Out[15]: OLS Regression Results
Dep. Variable: travel_time R-squared: 0.904
Model: OLS Adj. R-squared: 0.901
Method: Least Squares F-statistic: 32.88
Date: Fri, 08 Sep 2017 Prob (F-statistic): 0.000276
Time: 11:16:53 Log-Likelihood: -6.8398
No. Observations: 10 AIC: 19.68
Df Residuals: 7 BIC: 20.59
Df Model: 2
Covariance Type: nonrobust

coef std err t P>|t| [0.025 0.975]
Intercept 0.8687 0.052 -0.913 0.392 -3.119 1.381
x1 0.0011 0.010 0.102 0.000 0.038 0.065
n_no_of_deliveries 0.0234 0.021 4.176 0.004 0.401 1.446

Omnibus: 0.039 Durbin-Watson: 2.515
Prob(Omnibus): 0.981 Jarque-Bera (JB): 0.151
Skew: 0.074 Prob(JB): 0.827
Kurtosis: 2.419 Cond. No. 435
```

7

F value is 32, so this value 32.8 that that is 32.9 approximately then we can see the easy here the p value probability of stat this p value is less than 4.05. when you say alpha is equal to 5.5 % then also we have to reject the null hypothesis.

(Refer Slide Time: 05:00)

ANOVA table

Source	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	SSR	p	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$
Error	SSE	$n - p - 1$	$MSE = \frac{SSE}{n - p - 1}$	
Total	SST	$n - 1$		

8

So this is anova regression table what is the sources of error? Error due to regression variable error total regression sum of square error sum of square total sum of square look at the degrees of freedom that is more important. There are 1 independent variable the degrees of freedom is 1 for regression sum of square. For TSS that is the total sum of square for that $n - 1$, n is number of data set. So, when you subtract $n - 1 - p$ that is why we are getting $n - 1 - p$ is $n - p - 1$.

So, what is MSR is SSR divided by p, MSE is SSE divided by n - p - 1. So, F equal to MSR divided by MSE this is the F value which you got anova output, when we introduce both variable into the model corresponding anova table for that two independent variable regression model is this one.

So, this is another table first you find out regression sum of square error of sum of square SST, p is the degrees of freedom n is the number of independent variable. For SST the degree of freedom is n - 1. If you want to know the degrees of freedom for n - 1 - p that is n - p - 1 MSR =SSR divided by p. MSE = SSE divided by n - p - 1 finally we are getting F value. So, this F value is nothing but your 32.88.

(Refer Slide Time: 06:30)

t Test for individual significance

For any parameter β_i

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

TEST STATISTIC

$$t = \frac{b_i}{s_{b_i}}$$

REJECTION RULE

p-value approach: Reject H_0 if p-value $\leq \alpha$
 Critical value approach: Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$
 where $t_{\alpha/2}$ is based on a t distribution with $n - p - 1$ degrees of freedom.

Now will go for individual for each variable we will see that significant weather that individual variable is significant are not, for any parameter beta i H 0 equal to beta i equal to 0, H a beta i 0 equal to 0. The test statistics is bi divided by Sbi actually they should be this way bi – beta i divided by Sbi because we are assuming beta equal to 0 so the remaining is only bi divided by S bi this Sbi for ith independent variable we can see that is standard error.

When I go back you see this the standard error for x1 is this 0.010 the standard error for second independent variables 0.0221 that value we can get it this one. So for p value approach reject H 0 the p value is less than or equal to Alpha as usual. Critical value of approach reject H 0 t value is

below the lower tail or above the upper tail. Here the $t \alpha/2$ is based on the t-distribution with $n - p - 1$ degrees of freedom ok we will do that one.

(Refer Slide Time: 07:50)

t Test for individual significance

	$b_1 = .061135$	$s_{b_1} = .009888$
	$b_2 = .9234$	$s_{b_2} = .2211$

$t = .061135/.009888 = 6.18$
 $t = .9234/.2211 = 4.18$

So, beta 1 is 0.0611 where we got this one and going back, t test for individual significance from output of python model we get b1 equal 2.061135, b2 equal to 0.9234, sb1 equal to 0.00988, sb2 equal to 0.2211 where we are getting just see that b1. So, b1 is this value 0.0611 the Sb1 is this 0.011 I am going back, see that is b1 so 0.0098 so it can be 0.01 because after rounding ok. The b2 is 0.92, Sb2 is 0.221. So, what is the t formula of b 1 divided by Sb1 b1 is 0.061135 divided by 0.088, 6.8 here t is 4.18 this also you can verify.

Where this for first variable it is 6.18 see that 6.18, for second variable this is 4.16. Look at the p value, p value for first variables 0.000 2nd variable is 0.00 by looking at the p-value itself without to reject the null hypothesis. When we reject null hypothesis beta 1 not equal to 0 that mean there is a relation between x 1 and dependent variable y at the population level. Similarly for the second independent variable, also, we have to reject null hypothesis.

So, beta 2 not equal to 0 that means that there is a relation between x 2 that is the second variable number of deliveries and dependent variable. This is the way to interpret the Python output of this multiple regression model. And going to give a demo for the multiple regression model. Ok students we have seen the theory behind the multiple regression model. We have

come to Python background. I have already prepared the code as shown the output suppose. You want to do it demo in our class are you want to someone in this course.

(Video Start time: 10:15)

Go to this kernel option restart and clear output so what will get it there is a restart and clear all output when you do this way to see that only the quotes will be there, there would not be any output. Suppose you want to show to others what is going to be the output of this code you can do that way. So first we will import the necessary libraries import Pandas as pd from statsmodels dot formula dot api import ols, so this library is used for doing regression analysis.

As you know that the pandas is used for reading a loading the files. From stat model dot stat anova import anova_lm, so this library is used to see the output of anova table for regression model import matplotlib.pyplot as plt so this is used for plotting the figure. First I have stored my data set the file name is called tracking. I have loaded this data set into the object called df1 first will run the library then will see the what is the data set.

In this data set a when you look at this. The first one is the index column second one is the driving assignment third one is our independent variable that is the travel time. The next one is number of delivery third one is not the travel time. It is the distance travelled the x1 means distance travelled the next independent variable is number of deliveries. The travel time is our dependent variable here.

Here what were going to do? First, we are going to see what is the relation between x 1 and our dependent variable then we will see x 2 the number of deliveries versus travel time. Then we will see both independent variable together. Then will see what is the effect on the dependent variable? First will do the scatter plot then we can understand the relationship that trend between is independent variable and the dependent variable. This one is x1 is taken as a independent variable and y is taken as the different variable.

It seems to be there is some positive trend is there. Why this scatter plot is required if there is no relationship at all suppose the lines are in a horizontal manner that you need not do any regression analysis because there is no relation between this x and dependent variable. Now we

will take 2 variable then you got it. Now it is happening the green dot shows one variable and red dots shows another variable. Now, this is only second independent variable. Now will get the regression model the first regression model where we are going to consider only one independent variable. So that is model name is reg1 is equal to OLS().

The formula equal to travel underscore time is the dependent variable tilde symbol actually the first you to write the dependent variable tilde x1 in double quote data equal to df1. Because this tilde symbol even if you know the R programming, in R programming also similar Syntax will use that one. So, will you do this one you are getting the output of our regression model where only one independent variable is considered. So what is the first task is we have to construct the regression equation? What is regression equation y equal to $1.2739 + 0.0678 \times 1$.

Second one is where to locate the R square? R square is 0.664 if the R square is more than 50. It is considerably a good model even though it is 66 this is accepted. Then look at the F statistics, F statistics to 15.81 that look at the probability that is less than 0.05 as a whole model this regression model is acceptable. Remember that we are using only one independent variable within a look at the p value when the first independent variable 0.004 less than 0.05 when you look at an integer variable also this one variable is significant variable.

The next one what you are going to do? We are going to introduce both the independent variable together. Then we are going to see the impact on R square. So here I go to say that model is regression 2 where 2 is equal to OLS(formula equal to travel time tilde x1 plus I am adding second independent. If there is third independent variable plus you have to add the third independent variable. Then fit2 equal to reg2 fit.

So print fit 2.summary, look at this, first will come regression equation. Regression equation is y equal to $-0.8687 + 0.0611 \times 1 + 0.9234 \times 2$ otherwise number of deliveries. You compare the R square with the previously say R square is 0.664 now when you use to independent variable, the R square is increased. So, the goodness of it the model is increased when we introduce more independent variable. There is another term adjusted R square this is adjusted for number of independent variables.

When you are in keep on introducing more number of independent variables you have to monitor the value of R squared and adjusted R square what will happen when you introduce more independent variables R square will always will increase but adjust R square it will initially start increase after certain point it start decreasing that point You should stop adding more independent variable that means the R square value when is decreasing that the new variable which have introduced into the model is not helping to explain the regression model instead of it is going to disturb the existing model that is that a new variable it is the noise variable.

Look at the F value F is 32.8 look at the probability value it is less than 0.05 as a whole model so we are going to reject null hypothesis. So, what this F statistics says F statistics is used to test overall significance of the regression model that both x_1 x_2 by considering in this regression equation the model is valid. Then we will go for significance of each individual independent variable. There is x_1 when you look at the p-value it is 0.000 that means we have to reject the null hypothesis of beta 1 equal to 0.

When you rejected it that means beta 1 is not equal to 0 then you say beta 1 is not equal to 0, even if the population over there is a relation between x_1 and the y variable. Similarly look at the p value for the our second independent variable that is also less than 0.05 that means that the second variable also significant variable in our regression model. Sometime what will happen they may be different independent variables for some independent variable p value more than 0.05.

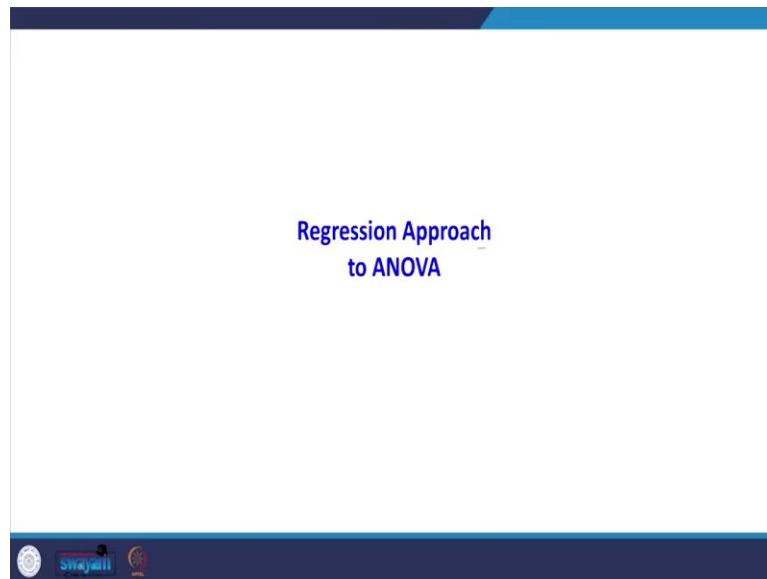
If it is more than 0.05 then you are writing regression model that corresponding independent variable has to be dropped. So meaning of the dropping is that with the help of sample data we can try regression equation by considering all independent variable but that cannot be generalized at the population level because certain variable cannot be significant at the population-level how to know that the variable is not significant we have to look at the p-value.

The p value is more than 0.05 we to accept null hypothesis, when we accept null hypothesis. that means beta 1 equal to 0 then there is no relation between that independent variable at the population. Other goodness of fit is we will see the what is the meaning of Durbin Watson in the

coming classes. Similarly there is one more measure to check the goodness of model AIC and BIC and we go for our Logistic regression, then I will explain what is the meaning and significance of AIC.

(Video End Time: 19:02)

(Refer Slide Time: 19:03)



So, far we have studied the regression analysis? Now with the help of regression I am going to tell you how to solve an anova problem. I have taken one sample problem that problem first time going to solve with the help of anova then with help of excel and good to solve it after that the same problem. I go to explain how to solve that anova problem with the help of regression analysis.

(Refer Slide Time: 19:34)

Regression Approach to ANOVA

- Three different assembly methods, referred to as methods A, B, and C, have been proposed.
- Managers at Chemitech want to determine which assembly method can produce the greatest number of filtration systems per week

A	B	C
58	58	48
64	69	57
55	71	59
66	64	47
67	68	49

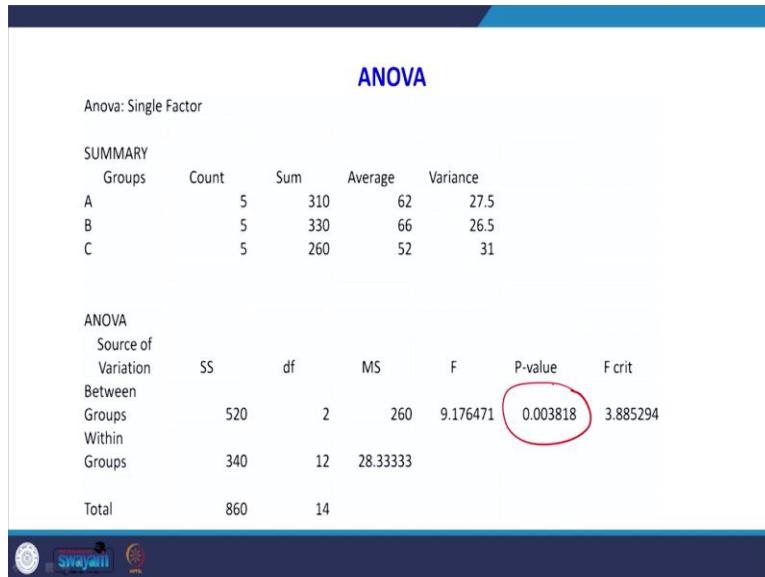


The problem is like this what is happening in the three column is there A, B and C. A Represent one type of assembly method B represents another type of assembly method C represents the third type of assembly methods. If you follow method A; 58, 64, 55, 66, 67 represents number of product which are assembled per week; similarly for B under the column B 58 69 71 64 68 represents number of product assemble to per week. Here the product is filtration system. As a manager I want to know which method is producing the better result.

That means if I follow method A or B or C which one will produce are will help me to assemble more number of products. This is a typical anova problem. So in anova what we used to do generally the null hypothesis, The null hypothesis $\mu_A = \mu_B = \mu_C$ that means the mean of the product assemble through method A equal to mean of the product assembled by method B equal to the product obtained by method C.

Obviously the alternative hypothesis $\mu_A \neq \mu_B \neq \mu_C$, the purpose of doing this is to identify which assembly method is more productive. In case if I accept null hypothesis all the three assembly methods are giving the same result I am not able to identify which method is better. In case if I reject my null hypothesis, I can clearly say which method is the better method which will give you more number of products assembled.

(Refer Slide Time: 21:44)



This I am go to solve with help of Excel enter the data in three column A column B column C. So go for data go for data analysis go for anova. Anova is a single factor because one way anova so the input range is I am selecting all this values, so labels in the First row yes, so when I say ok I am getting this output, what this mean is, when you look at the p-value here, it is 0.00382 it is less than 0.05. So I will be rejected my null hypothesis.

When I reject null hypothesis all the three assembly method not producing equal result, this was the output got it. How to interpret this output when you look at this p-value the p-value is less than 0.05 so I am rejected my null hypothesis.

(Refer Slide Time: 22:48)

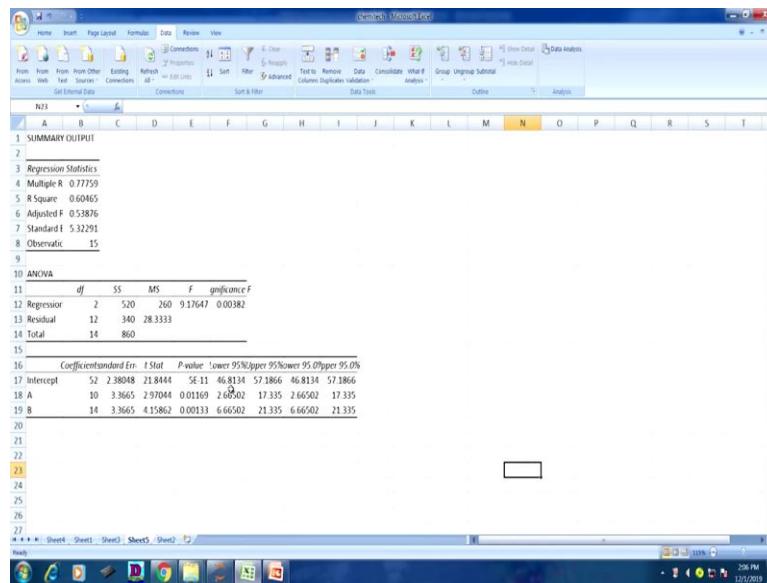
The table illustrates the mapping of assembly methods to binary variables (A and B). Method A is represented by column A (1 for method A, 0 for others) and column B (0 for method A, 1 for others). Method B is represented by column A (0 for method B, 1 for others) and column B (1 for method B, 0 for others). Method C is represented by column A (0 for method C, 1 for others) and column B (0 for method C, 1 for others).

		Dummy variables for the chemitech experiment	
		A	B
1	0	Observation is associated with assembly method A	
0	1	Observation is associated with assembly method B	
0	0	Observation is associated with assembly method C	

Now, this is what I have got in the previous slide I am going to get with the help of regression analysis. The regression analysis I am going to use the concept of dummy variable because there are three assembly methods is there. Generally 3 - 1 number of dummy variables is required. How I am creating dummy variable I am taking say A this is A is one dummy variable for example B is another variable.

If I save 1, 0 that represents the presence of 1 represents variable If you say 0,1 the presence of 1 represents on column B the represents the B variable see that the absence of 1 in both columns represent assembly method say that is why it is written here, see the 1,0 observation is associated with assembly line method A 0,1 represents the observation is associated with assembly method B 0,0 represents the observation is associated with assembly method C. So, I am going to do this modification then I am going to do regression analysis. So, with the help of regression I am going to explain here, but after sometime I go to explain to you. How to use Python. So, first now I will explain with help of Excel.

(Refer Slide Time: 24:05)



The screenshot shows an Excel spreadsheet titled "Untitled - Excel" with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	SUMMARY OUTPUT																			
2																				
3	Regression Statistics																			
4	Multiple R	0.77759																		
5	R Square	0.60465																		
6	Adjusted R	0.52876																		
7	Standard E	5.32291																		
8	Observatio	15																		
9																				
10	ANOVA																			
11		df	SS	MS	F	Significance F														
12	Regressor	2	520	260	9.17647	0.00382														
13	Residual	12	340	28.3333																
14	Total	14	860																	
15																				
16		Coefficients	standard Err.	t Stat	P value	'lower 95%	'upper 95%	'lower 95%	'upper 95%											
17	Intercept	52	2.38088	21.8444	5E-11	46.8134	57.1866	46.8134	57.1866											
18	A	10	3.3665	2.9704	0.01169	2.65902	17.335	2.65902	17.335											
19	B	14	3.3665	4.15862	0.00133	6.66502	21.335	6.66502	21.335											
20																				
21																				
22																				
23																				
24																				
25																				
26																				
27																				

This is the given data set. So what I have done this is my after coding. For example see that upto 58 to 67 this column up to this much it represents A so I have written 1,0,1,0,1,0,1,0 see that here up to 1,0,1,0 the presence of 1 represents assembly method A the absence of 1 represents a assembly method B. Similarly I have to type the remaining B values see that here 0,1,0,1,0,1,0,1. So this portion represents look at this the presence of 1 represents method B.

The last one is this portions I have taken 0,0 on both columns that means absence of one and both the method represents the Assembly method C. Now this one I going to do the regression analysis go for data analysis go for regression. Here the y value is this one x values here there are 2 dummy variable this one, when you run it you are getting this output you see that. When you look at the p-value here here also, you are getting 0.0038 that means here also you are rejecting the null hypothesis. I will explain how to interpret the coefficient of 52, 10, 14 in coming slides.

(Refer Slide Time: 25:52)

Dummy variables for the chemitech experiment

$E(y) = \text{Expected value of the number of units produced per week}$
 $= \beta_0 + \beta_1 A + \beta_2 B$

- If we are interested in the expected value of the number of units assembled per week for an employee who uses method C, our procedure for assigning numerical values to the dummy variables would result in setting $A = B = 0$.
- The multiple regression equation then reduces to

$$E(y) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$

Expected value of y is equal to expected value of number of units produced per week. This is the regression equation beta 0 + beta 1 A+ beta 2 B if you are interested in the expected value of the number of units assembled per week for an employee who uses method C our procedure for assigning numerical value to the dummy variable result in setting A equal to B equal to 0, suppose if you want to know the answer for assembly method C you to substitute A equal to 0 B equal to 0. When you substitute A equal to 0 B equal to 0 the expected value of y is nothing but your beta 0.

(Refer Slide Time: 26:35)

Dummy variables for the chemitech experiment

- For method A the values of the dummy variables are A = 1 and B = 0, and

$$E(y) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$

- For method B we set A = 0 and B = 1, and

$$E(y) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$



In case for method the value of the dummy variable is equal to 1 b equal to 0 when you substitute in the regression equation beta 0 + beta 1 because A values is 1 B value 0 so beta 0 + beta 1. If I want to know the expected value of assembly method B we have to set A equal to 0 and B equal to 1 when you substituted regression equation You are getting beta 0 + beta 2.

(Refer Slide Time: 27:07)

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.777593186							
R Square	0.604651163							
Adjusted R Square	0.53875969							
Standard Error	5.322906474							
Observations	15							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	2	520	260	9.176471	0.003818412			
Residual	12	340	28.33333					
Total	14	860						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	52	2.380476143	21.84437	4.97E-11	46.81338804	57.18661196	46.81338804	57.18661196
A	10	3.366501646	2.970443	0.011692	2.665023022	17.33497698	2.665023022	17.33497698
B	14	3.366501646	4.15862	0.001326	6.665023022	21.33497698	6.665023022	21.33497698



Now what we got it, when you look at this coefficients the intercepts is 52 that is your beta 0. This is beta 1 for coefficient of A this is beta 2 this is coefficient of B.

(Refer Slide Time: 27:23)

Estimation of $E(y)$

- $b_0 = 52$
- $b_1 = 10$
- $b_2 = 14$

Assembly Method	Estimation of $E(y)$
A	$b_0 + b_1 = 52 + 10 = 62$
B	$b_0 + b_2 = 52 + 14 = 66$
C	52



What will happen see beta 0 is 52 beta 1 equal to 10 B2 equal to 40 if you want know the estimated value of y for assembly method A you have to refer $b_0 + b_1$ how we got b_0 b_1 look at this beta 0 + beta 1 now beta 0 52, b 1 is 10 so totally 62. If you want to know the estimated value of y for assembly with B it is 52 + 14 how we got 52 + 14 using this equation beta 0 is 52 Beta 2 is 14, what is beta 2 look at this Beta 2 is 14. So when you substitute here we are getting 66. If you want to know the expected mean of methods C you have substitute A equal to 0 B equal to 0 you get only beta 0, beta 0 is estimate with help of b 0 that values 52 .

(Refer Slide Time: 28:24)

Testing the significance

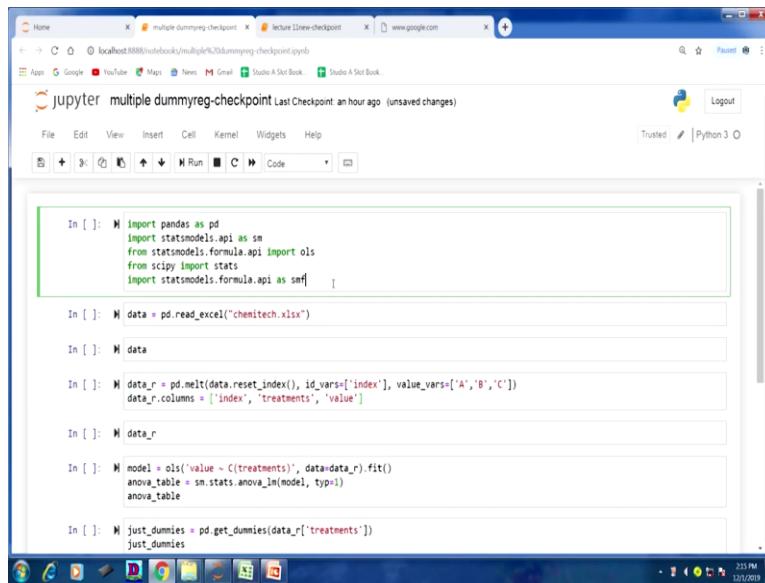
$$H_0: \beta_1 = \beta_2 = 0$$



Then we can go for significance test beta 1 equal to beta 2 equal to 0 this we have seen already we can go for t test or F test. In the F test if the value is less than 0.05 then we can say both the

variables are significant. What happened here when you do with the help of Excel you see that the p value of F see the p-value is less than 0.05 we can see the regression coefficient A and B is significant. So far I have done with the help of Excel. Now I am going to do the same problem python.

(Refer Slide Time: 29:01)



The screenshot shows a Jupyter Notebook window titled "jupyter multiple dummyreg-checkpoint". The code cell contains the following Python script:

```
In [ ]: import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols
from scipy import stats
import statsmodels.formula.api as smf

In [ ]: data = pd.read_excel("chemitech.xlsx")

In [ ]: data

In [ ]: data_r = pd.melt(data.reset_index(), id_vars=['index'], value_vars=['A','B','C'])
data_r.columns = ['index', 'treatments', 'value']

In [ ]: model = ols('value ~ C(treatments)', data=data_r).fit()
anova_table = sm.stats.anova_lm(model, typ=1)
anova_table

In [ ]: just_dummies = pd.get_dummies(data_r['treatments'])
just_dummies
```

First will import the necessary files like import Pandas as pd import statsmodels dot api dot sm from statsmodels dot com dot api import ols form scipy import stats import statsmodels dot formula dot api as smf the data is stored in a file called chemitech. So this is my data set so far this data set and going to do an anova after doing anova I go to check the result the same problem I run it with help of regression analysis.

For doing regression analysis I go to use the concept called dummy variable. So, first I will run this given data set anova so for that purpose. I am converting this data set into this form. What is that form? That I will show in the data_r, you see that all the treatments are in one column all the values are in one column model equal to wireless. Value is our dependent variable Tilde see the treatment is my dependent variable.

The command for regression analysis from the regression analysis I am going to get the anova table the anova table is this one? You see that the p-value 0.0038 when I am doing the same problem with the help of Excel also because the same result. So, what we are concluding here for

all the means or not equal so we are rejecting our null hypothesis. Now this problem we are going to do with the help of regression analysis by using the concept called dummy variables that one treatment that is assembly method with 3 variables ABC so that I am going to convert into dummy variables.

You look at this ABC is the presence of 1 represents A the presence of 1 here represents B the C column the presence of 1 represents C because in the treatment of three levels, we need only two dummy variables. So we are going to drop the column C then we are going to use the two column A and B. In excel also when I am solving meet you seen this kind of data. So what I am going to do I am going to drop this column C then I am going to add this dummy variables into the filename called step_1.

This one see that now the file is changed. Now the value is taken as it is only the column A and B is maintained. So for this dataset going to do the regression analysis, so the results is equal to smf dot ols step underscore 1 the value is my different variable sm dot add underscore constant step_1 A, B is my independent variable then I go to get the regression output. Now look at this regression output. Look at this probability that the p-value 0.00382 here also we are rejecting our null hypothesis.

Then look at the constant value constant is 52 that is b 0, b 1 is 10 and b 2 is 14 this value is taken for interpreting the output. Now look at the variable A and B both are the p values less than 0.05 both variables are significant the value of b0 b1 b2 can be used for interpreting as I explained in my slide for interpreting the output. In this class we have seen how to do the significance test for multiple regression model.

That significance test we have done with the help of two test one is F test and t test. F test is used to test the overall significance of the regression model, t test is used to check the individual significance of each independent variable. After that I have taken a sample problem. Then I explain the given a demo how to do the multiple regression, then we have interpreter the output of multiple regression model.

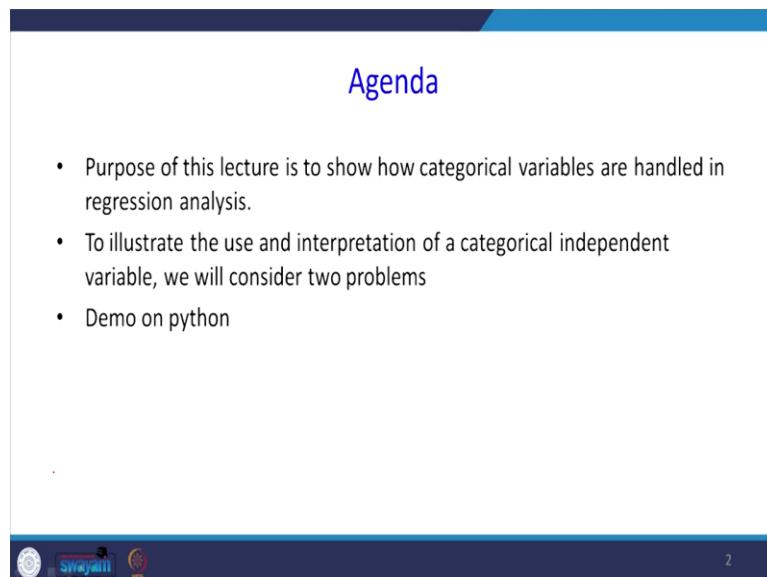
In the next class going to do another regression model that is where the independent variable that is categorical independent variable. So far what we have done that one dependent variable and independent variable both are continuous. There may be situation where that independent variable is categorical variable. For example gender is a categorical variable that can have only two option male or female in that case you to do some adjustment in existing; our regression model. How to do that one that you will see in the next class, thank you very much for listening.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 35
Categorical Variable Regression

Dear students, in this lecture, we will see how to handle Categorical Variable linear regression analysis. Whenever we do a linear regression analysis, the assumption is the nature of independent and dependent variable has to be continuous variable. Sometimes what will happen we have to include the categorical variable into independent variable category? How to handle that kind of regression analysis that we will see in this class?

(Refer Slide Time: 00:57)



The slide has a dark blue header and footer bar. The title 'Agenda' is centered in white font. The main content area is white with a dark blue border. A list of four bullet points is displayed:

- Purpose of this lecture is to show how categorical variables are handled in regression analysis.
- To illustrate the use and interpretation of a categorical independent variable, we will consider two problems
- Demo on python

In the footer bar, there are three small circular icons and the number '2'.

The Agenda of this lecturer is, to show how categorical variable are handled in regression analysis. Illustrate and will interpret how to do the categorical independent regression analysis. The same problem we will do in Python will explain how to code and how to do this categorical regression in Python programming.

(Refer Slide Time: 01:17)

What are dummy variables?

- Dummy variables, also called indicator variables allow us to include categorical data (like Gender) in regression models
- A dummy variable can take only 2 values, 0 (absence of a category) and 1 (presence of a category)

3

Another name for categorical variable is called dummy variable dummy variable also called indicator variable. It allows us to include categorical nature in regression analysis. For example, gender is one of a categorical data where there is only two levels are possible male or female. If dummy variable can take only two values, when it is gender category, for example, zero means absence of category and one means the presence of category. Here zero will be taken as their reference. With respect to zero, we will compare what will happen to another level of the categorical variable.

(Refer Slide Time: 01:51)

Example 1: Problem / Background

- Johnson Filtration, Inc., provides maintenance service for water-filtration systems.
- Customers contact Johnson with requests for maintenance service on their water-filtration systems
- To estimate the service time and the service cost, Johnson's managers want to predict the repair time necessary for each maintenance request
- Hence, repair time in hours is the dependent variable
- Repair time is believed to be related to two factors,
 - the number of months since the last maintenance service
 - the type of repair problem (mechanical or electrical).



Source: Statistics for Business & Economics, David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, Cengage Learning, 2013

4

We will take a problem with the help of problem I will explain how to use categorical variable into the regression analysis and how to interpret it. This problem is taken from statistics for

Business and Economics from David Anderson, Sweeney and Williams. It is Syncage Publication in 2003 to 2013 edition. Johnson filtration Incorporation provides maintenance service for water filtration systems.

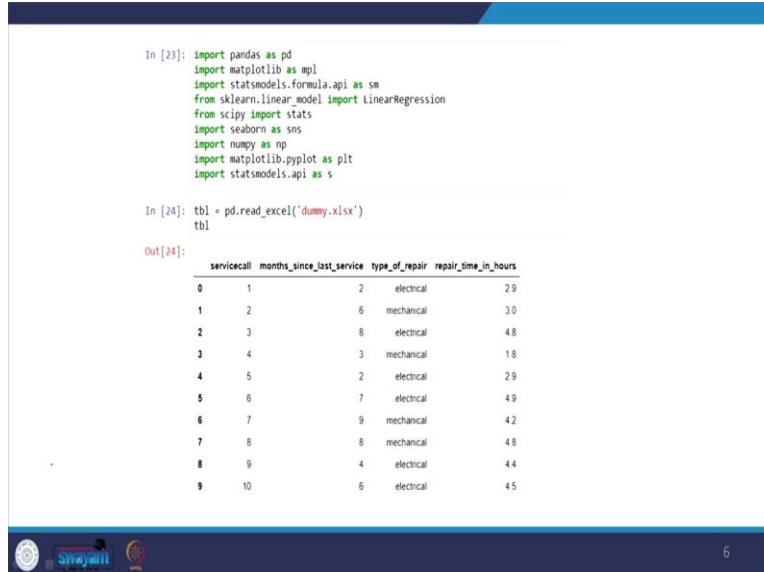
Customers contact Johnson's with a request for maintenance service on their water filtration system. To estimate the service time and the service cost Johnson's managers want to predict the repair time necessary for each maintenance request. Hence, the repair time in hours is the dependent variable. Repair time is believed to be related to two factors. One factor is number of months since the last maintenance service was done; second factor is the type of repair problem. Here the type of repair problem, mechanical or electrical is the categorical variable.

(Refer Slide Time: 02:53)

Data for the Johnson filtration example			
service call	months_since_last_service	type_of_repair	repair_time_in_hours
1	2	electrical	2.9
2	6	mechanical	3
3	8	electrical	4.8
4	3	mechanical	1.8
5	2	electrical	2.9
6	7	electrical	4.9
7	9	mechanical	4.2
8	8	mechanical	4.8
9	4	electrical	4.4
10	6	electrical	4.5

This is the given data. What is there is a Column 1 is the service call, the column 2 says months since the last service was done, in terms of month. Column 3 says the type of repair whether it is the repairs with respect to electrical system or mechanical system. The last column is repair time in hours. How much time it is taken for doing repairing?

(Refer Slide Time: 03:18)



```
In [23]: import pandas as pd
import matplotlib as mpl
import statsmodels.formula.api as sm
from sklearn.linear_model import LinearRegression
from scipy import stats
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as s

In [24]: tbl = pd.read_excel('dummy.xlsx')
tbl
```

	servicecall	months_since_last_service	type_of_repair	repair_time_in_hours
0	1	2	electrical	2.9
1	2	6	mechanical	3.0
2	3	8	electrical	4.8
3	4	3	mechanical	1.8
4	5	2	electrical	2.9
5	6	7	electrical	4.9
6	7	9	mechanical	4.2
7	8	8	mechanical	4.8
8	9	4	electrical	4.4
9	10	6	electrical	4.5

I have taken the screenshot of our python code get so I have to import necessary libraries like import Pandas as pd, import matplotlib as mpl, import statsmodels dot formula dot api as sm from sklearn.linear underscore model import LinearRegression from scipy import stats import seaborn sns, import numpy as np, import matplotlib.pyplot as plt, import statsmodels dot api as s. First we will load this regression file it is a data file I have saved in the name of dummy dotxlsx that we are going to save any object called tv1.

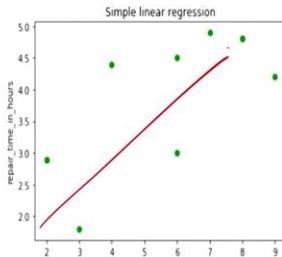
When you execute this one, we can see this is a data file. At the end of the class going to give the demo for this what are the codes which ever done in it. There also we can understand the steps. Here this is the data, display the data.

(Refer Slide Time: 04:10)

Linear Regression

```
In [41]: plt.scatter(tbl['months_since_last_service'],tbl['repair_time_in_hours'], color = "green")
plt.ylabel('repair_time_in_hours')
plt.title('Simple linear regression ')
```

```
Out[41]: Text(0.5,1,' Simple linear regression ')
```



7

But first will do the scatter plot between the months since last service and repair time in hours. When we look at this scatter plot, you see that there seems to be positive trend because when the month since last services more the repair time in hours also getting more. This is a simple linear regression considering only one independent variable. Here independent variable is continuous variable.

(Refer Slide Time: 04:41)

OLS Summary

```
In [44]: from statsmodels.formula.api import ols
Reg = ols(formula ~'repair_time_in_hours ~months_since_last_service', data =tbl)
Fit1 = Reg.fit()
print(Fit1.summary())
```

Dep. Variable:	repair_time_in_hours	R-squared:	0.534			
Model:	OLS	Adj. R-squared:	0.476			
Method:	Least Squares	F-statistic:	9.174			
Date:	Sat, 07 Sep 2019	Prob (F-statistic):	0.0163			
Time:	13:26:03	Log-Likelihood:	-10.692			
No. Observations:	10	AIC:	25.20			
DF Residuals:	8	BIC:	25.81			
DF model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025]	[0.975]
Intercept	2.1473	0.605	3.549	0.008	0.752	3.542
months_since_last_service	0.3041	0.100	3.029	0.026	0.073	0.536

$$y = 2.1473 + 0.3041 \text{ mmhs}$$



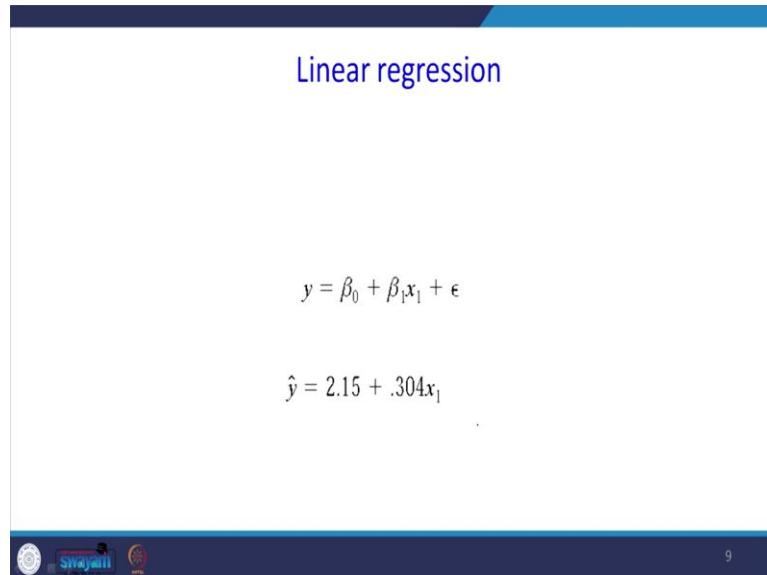
8

When you do the regression analysis, this is output of python. So, from statsmodels.formula .api import ols, ols is used for doing regression analysis. Here, the dependent variable is repair underscore time in hours tilde sign independent variables months since last service. When you look at this series, y intercept, I can write y equal to 2.1473 + 0.3041 because this is independent

variables months since the last service was done. Look at the R square. R square is 53.4 % look at the P value of this independent variable here. Here, it is significant because it is less than 0.05.

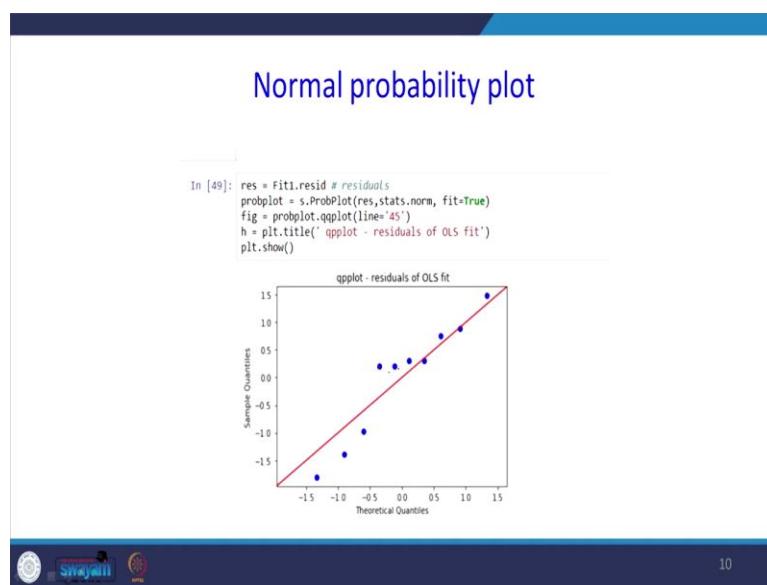
Now what we are going to do residual plots for this problem?

(Refer Slide Time: 05:38)



You look at this is $2.15 + 0.304 X_1$ is our regression model.

(Refer Slide Time: 05:47)



When we use this to do regression model, When you do the normal probability plot, look at this it has to all the probability points has to align with is red point. What is happening is there are so many points it is away from the red line. So, we can say that even if the, the residual plot is not appropriate, so the data, the error is not following normal distribution.

(Refer Slide Time: 06:10)

Creating dummies

```
In [34]: just_dummies = pd.get_dummies(tbl['type_of_repair'])
just_dummies
```

Out[34]:

	electrical	mechanical
0	1	0
1	0	1
2	1	0
3	0	1
4	1	0
5	1	0
6	0	1
7	0	1
8	1	0
9	1	0

$y = a + b_1 x_1 + b_2 x_2$

$y = a + b_1 x_1 + b_2(1)$

$y = a + b_1 x_1 + b_2(0)$

11

First, we will create a dummy variable for the categorical data. How to create a dummy variable for this categorical data? so that new dummy variable I going to call it is just underscore dummies equal to pd.get_dummies where the filename which column has to be converted into Dummies. So, the type of repair that is the value where we have written, whether the problem is related to mechanical or electrical.

So, when we display the just dummies see that that one variable is know, it is taken into 2 parts. 1 is for Electrical so, the presence of one says electrical; the absence of one says mechanical. There are 2 columns is there which is the dummy variable. So what happened both are same whether we can use this variable interval into our new regression model or this variable for Our new regression model, if you take electrical equal to 1.

So the equation be written as y equal to $a + b_1 x_1 + b_2 x_2$. Here, x_1 is independent variable. The b_2 value will be 1 if it is suppose we write if the problem the, this is the common regression equation. In this regression equation, when you substitute x_2 equal to 1 that equation for Electrical problem related to electrical repair $a + b_1 x_1 + b_2(1)$ this equation for repair due to electrical problem. Instead of this y equal to $a + b_1 x_1 + b_2(0)$ this is what problem related to mechanical. You can reverse also, no problem.

Mechanical can be taken as 1 and electrical can be taken as zero. There will not be problem in the interpretation.

(Refer Slide Time: 08:06)

DATA FOR THE JOHNSON FILTRATION EXAMPLE WITH TYPE OF REPAIR INDICATED BY A DUMMY VARIABLE ($x_2 = 0$ FOR MECHANICAL; $x_2 = 1$ FOR ELECTRICAL)			
Customer	Months Since Last Service (x_1)	Type of Repair (x_2)	Repair Time in Hours (y)
1	2	1	2.9
2	6	0	3.0
3	8	1	4.8
4	3	0	1.8
5	2	1	2.9
6	7	1	4.9
7	9	0	4.2
8	8	0	4.8
9	4	1	4.4
10	6	1	4.5

12

This was the data which we have converted into dummy variable. Month since last service 1 represents problem related to electrical Zero represents problem related to mechanical. This was our Y is our dependent variable.

(Refer Slide Time: 08:22)

Adding dummies to table																																																									
<pre>In [38]: just_dummies = pd.get_dummies(tbl[['type_of_repair']]) step_1 = pd.concat([tbl, just_dummies], axis=1) step_1 step_1.drop(['type_of_repair', 'mechanical'], inplace=True, axis=1) # to run the regression we want to get rid of the strings 'mechanical' and 'electrical' # and we want to get rid of one dummy variable to avoid the dummy variable trap # arbitrarily chose 'mechanical', coefficients on "electrical" would show effect of "electrical" # relative to "mechanical"</pre>																																																									
<pre>In [39]: step_1</pre>																																																									
<pre>Out[39]:</pre>																																																									
<table border="1"> <thead> <tr> <th>servicecall</th> <th>months_since_last_service</th> <th>repair_time_in_hours</th> <th>electrical</th> </tr> </thead> <tbody> <tr> <td>0</td> <td>1</td> <td>2</td> <td>2.9</td> <td>1</td> </tr> <tr> <td>1</td> <td>2</td> <td>6</td> <td>3.0</td> <td>0</td> </tr> <tr> <td>2</td> <td>3</td> <td>8</td> <td>4.8</td> <td>1</td> </tr> <tr> <td>3</td> <td>4</td> <td>3</td> <td>1.8</td> <td>0</td> </tr> <tr> <td>4</td> <td>5</td> <td>2</td> <td>2.9</td> <td>1</td> </tr> <tr> <td>5</td> <td>6</td> <td>7</td> <td>4.9</td> <td>1</td> </tr> <tr> <td>6</td> <td>7</td> <td>9</td> <td>4.2</td> <td>0</td> </tr> <tr> <td>7</td> <td>8</td> <td>8</td> <td>4.8</td> <td>0</td> </tr> <tr> <td>8</td> <td>9</td> <td>4</td> <td>4.4</td> <td>1</td> </tr> <tr> <td>9</td> <td>10</td> <td>6</td> <td>4.5</td> <td>1</td> </tr> </tbody> </table>				servicecall	months_since_last_service	repair_time_in_hours	electrical	0	1	2	2.9	1	1	2	6	3.0	0	2	3	8	4.8	1	3	4	3	1.8	0	4	5	2	2.9	1	5	6	7	4.9	1	6	7	9	4.2	0	7	8	8	4.8	0	8	9	4	4.4	1	9	10	6	4.5	1
servicecall	months_since_last_service	repair_time_in_hours	electrical																																																						
0	1	2	2.9	1																																																					
1	2	6	3.0	0																																																					
2	3	8	4.8	1																																																					
3	4	3	1.8	0																																																					
4	5	2	2.9	1																																																					
5	6	7	4.9	1																																																					
6	7	9	4.2	0																																																					
7	8	8	4.8	0																																																					
8	9	4	4.4	1																																																					
9	10	6	4.5	1																																																					

13

When you do the regression analysis, see that just_dummy is pd.get underscore dummies p1 is a type of repair. So here what I have done? I have displayed, I have dropped the certain columns what column I have dropped, I have dropped the column that is type of repair then I have added

only dummy variable with respect to the electrical repair. That is why this column has come. So, now this is going to this is the last column that is under electrical heading.

It is going to be taken as independent variable that will do the regression analysis.

(Refer Slide Time: 08:54)

```
In [20]: result = sm.OLS(step_1['repair_time_in_hours'], s.add_constant(step_1[['months_since_last_service', 'electrical']])).fit()
print(result.summary())

```

OLS Regression Results

Dep. Variable:	repair_time_in_hours	R-squared:	0.859			
Model:	OLS	Adj. R-squared:	0.819			
Method:	Least Squares	F-statistic:	21.36			
Date:	Sat, 07 Sep 2019	Prob (F-statistic):	0.0005			
Time:	13:08:09	Log-likelihood:	-4.6200			
No. Observations:	10	AIC:	15.24			
Df Residuals:	7	BIC:	16.15			
Df Model:	2					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	0.9305	0.467	1.993	0.087	0.174	2.035
months_since_last_service	0.3876	0.063	6.195	0.000	0.240	0.536
electrical	1.2627	0.314	4.020	0.000	0.520	0.005

Omnibus: 3.357 Durbin-Watson: 1.136
 Prob(Omnibus): 0.187 Jarque-Bera (JB): 1.663
 Skew: 0.994 Prob(JB): 0.435
 Kurtosis: 2.795 Cond. no. 22.0

y = 0.9305 + 0.3876
 months_since_last_service + electrical
 1.2627

Result equal to sm.OLS(step_one[‘repair_time_in_hours’] is taken as a dependent variable. Months underscore since underscore last underscore service taken as independent variable. So this electrical is taken as reference because that column where 1 means electric repair 0 means mechanical repair. When you look at this, you see this equation can be written as Y equal to 0.9305 Plus months since last underscore service ,See coefficient for this one is 0.3876 + electrical 1.2627.

So look at R square it is 0.85 previously, the R square was when there is only one independent variable I m going back previously asked for R is only for 0.534 when we introduce another variable what has happened, the R square is increased to 0.859. So, F statistics corresponding probability the p-value is very low 0.005 so as a whole this regression model is significant. When we look at the individual independent variable, for example, months_since there is independent variable 1, the P value is less than 0.01.

So we can say this variable is significant. Similarly, for the second one the type of repair, where electrical is taken as the reference this also less than 0.05, so, this is also a significant variable.

(Refer Slide Time: 10:42)

Dummy regression

$$x_2 = \begin{cases} 0 & \text{if the type of repair is mechanical} \\ 1 & \text{if the type of repair is electrical} \end{cases}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$\hat{y} = .93 + .388x_1 + 1.26x_2$$

$x_2 = 1 - \text{Electrical}$

$x_2 = 0 - \text{Mechanical}$

15

Now, this is the regression equation \hat{y} equal to $0.93 + 0.388 X 1 + 1.26 x_2$. If x_2 equal to 1 means electrical if I say x_2 equal to one it is related problem related to electrical if x_2 be 0 it is related to mechanical.

(Refer Slide Time: 11:15)

Interpreting the Parameters

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 \cancel{x_2}$$

$$E(y \mid \text{mechanical}) = \beta_0 + \beta_1 x_1 + (\beta_2(0)) = \beta_0 + \beta_1 x_1 \quad \text{Equation 1}$$

$$\begin{aligned} E(y \mid \text{electrical}) &= \beta_0 + \beta_1 x_1 + \beta_2(1) = \beta_0 + \beta_1 x_1 + \beta_2 \\ &= (\beta_0 + \beta_2) + \beta_1 x_1 \end{aligned} \quad \text{Equation 2}$$

16

The most important part, that is, interpreting the parameters. We know the expected value of Y equal to beta 0 + beta 1 x 1 + beta 2 x2 when you substitute equal to 1, when you substitute this x 2 = 0 that equation for mechanical, problem related to mechanical. So, beta 0 + beta 1 x1 beta x2 0 so this term will become there Beta 0 + beta 1 X 1 will be there. When substitute this x 2 equal to one that equations for the problem related to electrical.

So, E expected value y electrical equal to beta 0 + beta 1 x1 so, beta 2(1) what is happening so, beta 0 beta 2 that can be grouped that this will be beta one x1. See, both equations are same, both equation having the same slope Beta 1 only it differs by this extra value in our Y intercept how much with Beta 2.

(Refer Slide Time: 12:16)

Interpreting the Parameters

- Comparing equations, we see that the mean repair time is a linear function of x_1 for both mechanical and electrical repairs.
- The slope of both equations is β_1 , but the y -intercept differs.
- The y -intercept is β_0 in equation 1 for mechanical repairs and $(\beta_0 + \beta_2)$ in equation 2 for electrical repairs.

Comparing equations 1 and 2 we see that the mean repair time is linear function of X_1 for both mechanical and electrical repair. The slope of both equation is beta 1, but the y-intercept differs. The y intercept is beta 0 in equation 1 for mechanical repairs and beta 0 + beta 2 in equation 2 for Electrical repairs.

(Refer Slide Time: 12:45)

Interpreting the Parameters

- The interpretation of β_2 is that it indicates the difference between the mean repair time for an electrical repair and the mean repair time for a mechanical repair.
- If β_2 is positive, the mean repair time for an electrical repair will be greater than that for a mechanical repair; if β_2 is negative, the mean repair time for an electrical repair will be less than that for a mechanical repair.
- Finally, if $\beta_2 = 0$, there is no difference in the mean repair time between electrical and mechanical repairs and the type of repair is not related to the repair time.



18

The interpretation of Beta 2 is that it indicates the difference between the mean repair time of electrical repair and the mean repair time of mechanical repair. So the time differs by with this unit of this Beta 2. Beta 2 is positive the mean repair time for electrical repair will be greater than that of the mechanical repair. In our problem, if beta 2 is positive, if the beta 2 is negative the mean repair time for an electrical repair will be less than that of mechanical repair. If finally you Beta 2 equal to zero there is no difference in the mean repair time between electrical and mechanical repairs.

And the type of repair is not related to repair time. This is most important because after doing a dummy variable regression you have to interpret it. The interpretation is this way. The first thing is you have to look at what is the sign of this Beta 2. Beta 2 is positive or negative. Then in case the beta 2 is 0, we can save the type of the time taken to repair that filter is nothing to do with the type of problem it has occurred. Whether it is problems related to mechanical repair or problem related to electrical repair.

(Refer Slide Time: 14:01)

Interpreting the Parameters

- In effect, the use of a dummy variable for type of repair provides two estimated regression equations that can be used to predict the repair time, one corresponding to mechanical repairs and one corresponding to electrical repairs.
- In addition, with $\beta_2 = 1.26$, we learn that, on average, electrical repairs require 1.26 hours longer than mechanical repairs.

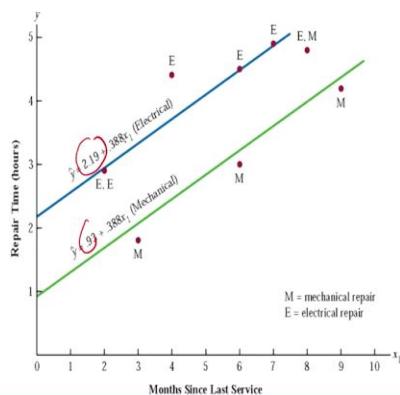
19

In effect, the use of dummy variable for type of repair provides 2 estimated regression equation that can be used to predict the repair time, one corresponding to mechanical repair and another corresponding to electrical repairs, in addition, beta 2 = 1.26 we are getting this 1.26, going back, this 1.26. This 1.26 we learnt that the average electrical repairs required 1.26 longer than the mechanical repairs because for electrical repairs we have taken $x_1 = 1$, for mechanical repair, we have taken $x_1 = 0$.

So, the electrical repair is taken as the reference. What is the meaning of that is that the 1.26 time units the electrical repair is taking longer time than mechanical repairs. Look at this picture.

(Refer Slide Time: 14:55)

Interpreting the Parameters



20

The green one is for mechanical repair when substitute $x_2 = 0$ here, the blue one is for electrical repair, very extreme cold one. Look at this one. 2.19, this is 0.19. Both the slopes are same. This slope is 0.388 for this equation and this equation. Only the intercept is differs.

(Refer Slide Time: 15:13)

More Complex Categorical Variables

- A categorical variable with k levels must be modeled using $k - 1$ dummy variables.
- Care must be taken in defining and interpreting the dummy variables.



21

What is the logic is that here we have we have seen only two levels. Sometimes, there may be more than two levels. So, the number of a categorical variable with k levels must be modeled using $k-1$ dummy variable. What happened previously there was a 2 level, so we have taken only one dummy variable x_2 . So there are three levels you have to take $3 - 1$ that is a 2 dummy variable. Care must be taken in defining and interpreting the dummy variable.

What is the care here is what is the value we have assigned is equal to 1. For example, electrical repair, you take an equal to one that equation is integrated with respect to $x_2 = 1$.

(Refer Slide Time: 15:54)

Example 2: Problem / Background

- The manager of a small sales force wants to know whether average monthly salary is different for males and females in the sales force.
- He obtains data on monthly salary and experience (in months) for each of the 9 employees as shown on the next slide.



We will go for another problem. This problem is taken from statistics for management from Lemen N Rubeen. The manager of a small sales force wants to know whether the average monthly salary is different for males and females in a sales force. He obtained a data on monthly salary and experience for each of 9 employees as shown in the next slide.

(Refer Slide Time: 16:20)

Data

Employee	Salary	Gender	Experience
1	7.5	Male	6
2	8.6	Male	10
3	9.1	Male	12
4	10.3	Male	18
5	13	Male	30
6	6.2	Female	5
7	8.7	Female	13
8	9.4	Female	15
9	9.8	Female	21

Look at this. This is there are nine employees their salary, there is gender, there is experience. Now what you are going to do in this example, what is the salary of the females even though they have equal experience with the male, whether females are discriminated or not when we can say that they are getting discriminated, even though they have equal experience with male, they are getting lesser salary that means the females are discriminated.

(Refer Slide Time: 16:49)

```
In [50]: tbl2 = pd.read_excel('dummy2.xlsx')
tbl2
```

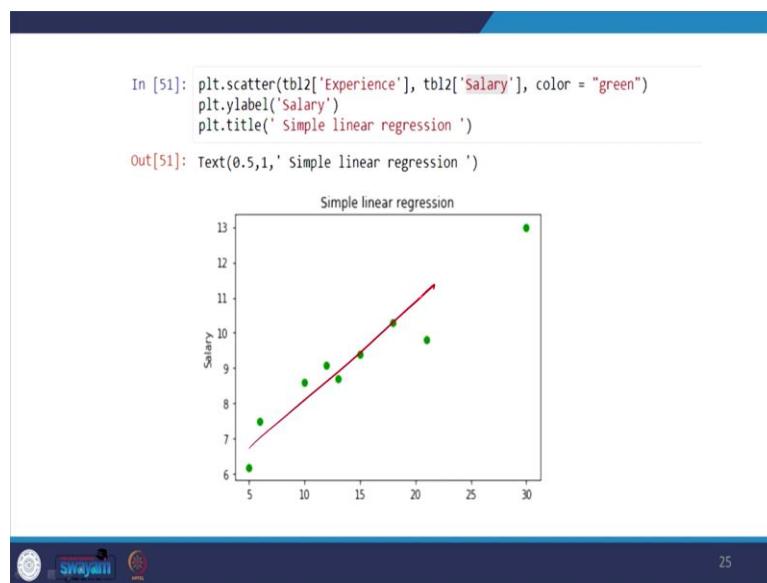
Out[50]:

	Employee	Salary	Gender	Experience
0	1	7.5	Male	6
1	2	8.6	Male	10
2	3	9.1	Male	12
3	4	10.3	Male	18
4	5	13.0	Male	30
5	6	6.2	Female	5
6	7	8.7	Female	13
7	8	9.4	Female	15
8	9	9.8	Female	21

24

First, we will import the data. Here are imported in the object called `tbl2 = pd.read_excel`. The excel data where I have stored this problem is in the filename called `dummy2`. So, when I show this. Look at this, this is the employee salary, gender and experience. Next, what we are going to do?

(Refer Slide Time: 17:14)



25

We are going to find out the scatter plot or is there any trend between the experience and the salary? It seems to be there is a positive trend. But look at the residual plot. What is this equation?

(Refer Slide Time: 17:29)

```
In [59]: Reg2 = ols(formula ="Salary ~ Experience", data =tbl2)
Fit2 = Reg2.fit()
print(Fit2.summary())

OLS Regression Results
=====
Dep. Variable: Salary R-squared: 0.926
Model: OLS Adj. R-squared: 0.915
Method: Least Squares F-statistic: 87.61
Date: Sat, 07 Sep 2019 Prob (F-statistic): 3.10e-05
Time: 14:18:45 Log-likelihood: -6.2491
No. Observations: 9 AIC: 16.50
Df Residuals: 7 BIC: 16.89
Df Model: 1
Covariance Type: nonrobust
=====
coef std err t P>|t| [0.025 0.975]
-----
Intercept 5.8093 0.404 14.386 0.000 4.854 6.764
Experience 0.2332 0.025 9.360 0.000 0.174 0.292
-----
Omnibus: 2.443 Durbin-Watson: 1.171
Prob(Omnibus): 0.295 Jarque-Bera (JB): 1.432
Skew: -0.918 Prob(JB): 0.489
Kurtosis: 2.331 Cond. No. 35.8
=====
```

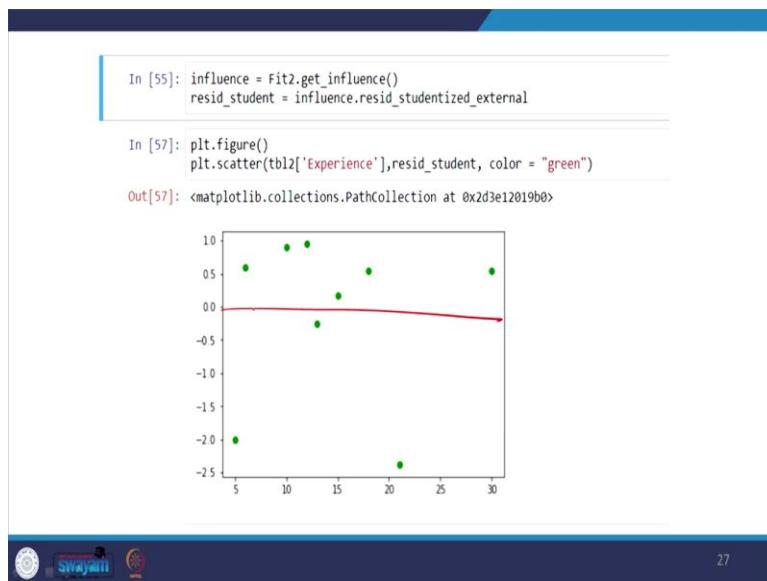
$$Y = 5.8 + 0.2332 \text{Exp}$$



26

Y equal to see, R square is 0.926. See, the experience is the independent variable. Experience is the because p value is less than 0.05 we can say as the significant value. So we can write Y equal to $5.8 + 0.2332$ experience. This is a regression equation. Ok now let us do the residual plot. For this we will do the error analysis.

(Refer Slide Time: 17:58)

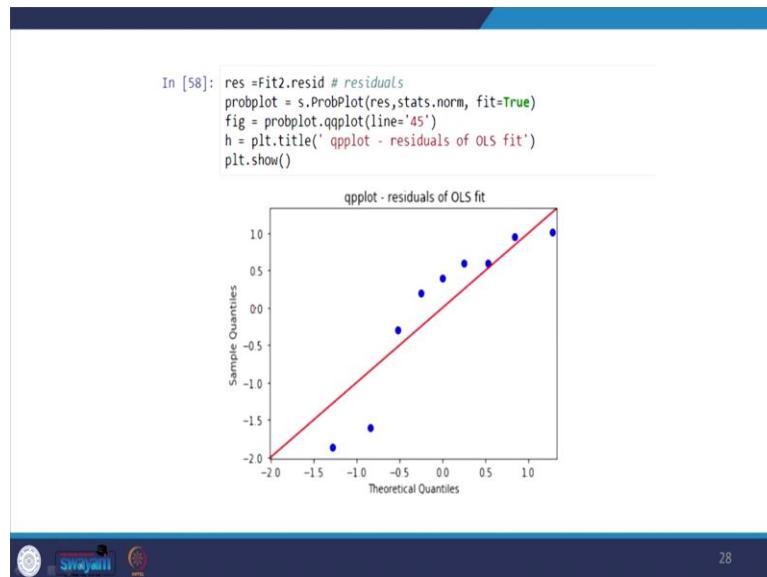


27

We will do the Residual analysis. You see that most of the points support, taken as the reference. This is a standardized residuals. Most of the points are you should be randomly it has to be distributed. Most of the points are above this way, there is a zero line. That means there is a

problem in assumption. Otherwise there might be some other variable that may affect the salary apart from experience.

(Refer Slide Time: 18:25)



Look at the see that quantile plot. You see that here also most of the points are above the pointers it has to sit on this red line, but it is not sitting on red Line then there is a problem in the assumption of that equal variance. That means error is not following equal variance.

(Refer Slide Time: 18:42)

Creating a dummy variable for gender

- Categorical data is included in regression analysis by using dummy variables
- For example, we can assign a value of 0 for males and 1 for females in our data so that a MR model can be developed

$\chi_1 = 0 \text{ males}$

$\chi_2 = 1 \text{ females}$

Employee	Salary	Gender
1	7.5	0
2	8.6	0
3	9.1	0
4	10.3	0
5	13	0
6	6.2	1
7	8.7	1
8	9.4	1
9	9.8	1

Swayam

Now, what we have done in this data. Categorical data is included in the regression analysis by using dummy variable here what you have done? Zero for males, 1for females. What has taken

zero also as reference or one also reference? So, one for male, female is taken as a reference now data, so that a multiple regression model can be developed. We will do that one.

(Refer Slide Time: 19:08)

In [24]: just_dummies2 = pd.get_dummies(tbl2['Gender'])
just_dummies2

Out[24]:

	Female	Male
0	0	1
1	0	1
2	0	1
3	0	1
4	0	1
5	1	0
6	1	0
7	1	0
8	1	0

From the given data, I have converted into dummy variable, one dummy variable for female because there are two level female and male. So, male is taken as one female taken is zero. The coding is that zero is taken as male one is taken female. So in this we are going to take this column for our further analysis. So, how to interpret this 0 means female one means male.

In creating a dummy variable for gender, we are going to follow this notation $x_2 = 0$ means male x_2 equal to 1 is taken as a female. So, after creating dummy variable first how to create a dummy variable in Python just _ dummies to that is a variable which I have given, pd.get_dummies. This was the command for making dummy variable. So we are going to take female column for further analysis. Zero means male one means female.

(Refer Slide Time: 20:10)

```

In [62]: step_1 = pd.concat([tbl2, just_dummies2], axis=1)
step_1.drop(['Gender', 'Male'], inplace=True, axis=1)
# to run the regression we want to get rid of the strings 'male' and 'female'
# and we want to get rid of one dummy variable to avoid the dummy variable trap
# arbitrarily chose 'male', coefficients on 'female' would show effect of 'female'
# relative to 'male'

result = sm.OLS(step_1['Salary'], s.add_constant(step_1[['Female']]))

print(result.summary())

```

OLS Regression Results

Dep. Variable:	Salary	R-squared:	0.107			
Model:	OLS	Adj. R-squared:	-0.020			
Method:	Least Squares	F-statistic:	0.8426			
Date:	Sat, 07 Sep 2019	Prob (F-statistic):	0.389			
Time:	14:23:57	Log-Likelihood:	-17.455			
No. Observations:	9	AIC:	38.91			
Df Residuals:	7	BIC:	39.38			
Df Model:	1					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
const	9.7000	0.853	11.367	0.000	7.682	11.718
Female	-1.1750	1.280	-0.918	0.389	-4.202	1.852

Omnibus: 0.387 Durbin-Watson: 1.912
 Prob(Omnibus): 0.824 Jarque-Bera (JB): 0.280
 Skew: 0.330 Prob(JB): 0.869
 Kurtosis: 2.441 Cond. No. 2.51

$y = 9.7 - 1.1750 x_1$

31

This was our Python output for that regression analysis. When you look at this, R square is 0.107 but look at here, first I will write the regression equation. $Y = 9.7 - 1.1750 x_1$. How to interpret this result you see that in the x_1 is not the significant value here it is not significant. At the sample data level, what is the meaning of x_1 ? Look at this. If you write $x_2=1$ here x_1 equal to 1 is not x_2 it is x_1 .

When you substitute $x_1 = 1$ this one, this coefficient says, it is negative. What is the meaning of these negative is this female is getting lesser salary when compared to male by this much unit because it is a negative sign, we go for interpretation.

(Refer Slide Time: 21:23)

More on the intercept and slope

- The value of the intercept, 9.70, is the average salary for males (as we coded gender=1 for females and 0 for males)
- The value of the slope, -1.175, tells us that the average females salary is lower than the average male salary by 1.175.

32

The value of the intercept is 9.7 the average salary for males has been coded a gender 1 for female ok then, 1 for female and 0 for males. So, the value of the slope is - 1.175 tells us that the average salary is lower than the average male salary by 1.175. What is the meaning of this? Females are getting 1.175 units they are getting lesser salary when compared to male. If it is a positive then we can interpret that. When compared to male females are getting more salary because the negative you are saying that when compared to male, females are getting less salary.

(Refer Slide Time: 22:09)

```
In [25]: step_1 = pd.concat([tbl2, just_dummies2], axis=1)
step_1.drop(['Gender', 'Male'], inplace=True, axis=1)
# to run the regression we want to get rid of the strings "male" and "female"
# and we want to get rid of one dummy variable to avoid the dummy variable trap
# arbitrarily chose "male", coefficients on "female" would show effect of "female"
# relative to "male"

result = sm.OLS(step_1['Salary'], s.add_constant(step_1[['Experience', 'Female']])).fit()
print(result.summary())

OLS Regression Results
-----
Dep. Variable: Salary R-squared: 0.974
Model: OLS Adj. R-squared: 0.965
Method: Least Squares F-statistic: 111.6
Date: Sat, 07 Sep 2019 Prob (F-statistic): 1.80e-65
Time: 12:33:40 Log-Likelihood: 1.5742
No. Observations: 9 AIC: 9.150
Df Residuals: 6 BIC: 9.742
Df Model: 2
Covariance Type: nonrobust
-----
            coef std err      t   P>|t|    [0.025  0.975]
-----
const    6.2485  0.291  21.439  0.000   5.535   6.962
Experience  0.2271  0.016  14.089  0.000   0.188   0.267
Female    -0.7890  0.218  -3.599  0.019  -1.172  -0.306
-----
Omnibus:        0.110 Durbin-Watson:   2.181
Prob(Omnibus):  0.947 Jarque-Bera (JB):  0.198
Skew:          0.174 Prob(JB):    0.906
Kurtosis:       2.363 Cond. No.     44.8
-----
```

$y = 6.2485 + 0.2271 \text{Experience} - 0.7890 \text{Female}$

Now what we are going to do? We are going to introduce the previously we considered only gender. That is a female is taken as a reference. Now, we are going to introduce the experience also. When you introduce the experience also know the regression equation is y equal to 6.2485 + 0.2271, experience - 0.7890 female. Now look at the p-value these p-values now less than 0.05. Now, here the gender is significant variable.

In your previous slide, when you go back in this slide, you see that the p value is not significant. So we cannot say there is a gender discrimination. We can write a regression equation with the help of sample data, but at the population level, there is no connection between Gender and their salary because the relation between x_1 that is there gender and the salary there is no relationship. That means here the both female and male are getting same salary.

But when we introduce our experience was one of the variable now the general also is significant, so by considering experience and the gender, now gender also one of their significant because you are the P value is less than 0.05. We look at the f value the f value is very low. The probability value also very low as a whole model, this model is significant individually also all the variables are significant.

(Refer Slide Time: 23:47)

```

In [63]: step_1 = pd.concat([tbl2, just_dummies2], axis=1)
step_1.drop(['Gender', 'Female'], inplace=True, axis=1)

result = sm.OLS(step_1['Salary'], s.add_constant(step_1[['Male']])).fit()
print(result.summary())

```

OLS Regression Results

Dep. Variable:	Salary	R-squared:	0.107			
Model:	OLS	Adj. R-squared:	-0.020			
Method:	Least Squares	F-statistic:	0.8426			
Date:	Sat, 07 Sep 2019	Prob (F-statistic):	0.389			
Time:	14:27:56	Log-likelihood:	-17.455			
No. Observations:	9	AIC:	38.91			
Df Residuals:	7	BIC:	39.30			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	p> t	[0.025	0.975]
const	8.3250	0.954	8.735	0.000	6.269	10.781
male	1.1750	1.280	0.918	0.389	-1.852	4.202

General Statistics

Count:	0.397	Bartlett's test:	0.12
Prob (Chi-sq):	0.424	Jarque-Bera (JB):	0.389
Skew:	0.310	Prob (JB):	0.869
Kurtosis:	2.441	Cond. No.:	2.77

34

What would happen if we used zero for females and one for males in our data. Would our results be any different right? So for that purpose we have done some modification here. For example, gender female it is just reversed. You see that there is a difference in intercept but the slope is same but the slope sign is different. So what is the meaning here? The male right, because the male is 1, the males are getting 1.175 unit of higher salary when compared to females.

(Refer Slide Time: 24:27)

Male = 1, female = 0

- Not really – With coding as above, the intercept would change to 8.525 (the average female salary), the slope for gender would still be 1.175, but now it would have a positive sign (reflecting that average male salary is higher than average female salary by 1.175).
Predicted salaries from the model for males / females would not change no matter how dummy variable is coded

35

So what happened is there any difference in the result not really. With the coding as above, the interested change to 8.525 see that 8.525 the slope of the gender would still 1.175, but it would have a positive sign reflecting that the average male salary is higher than average female salary by 1.175. So predicted salaries from the model for males and females would not change no matter how the dummy variable is coded.

(Refer Slide Time: 25:00)

More on dummy variables

- For gender, we had only 2 categories – female and male – thus we used a single 0/1 variable for this
- When there are more than 2 categories, the number of dummy variables that should be used equals the number of categories minus 1
- No. of Dummy Variables = No. of levels -1

36

Sometimes, what will happen, they may be more than one dummy variable, how that in our problem. We have only two levels, sometimes there are three levels. We should have to Dummy variable will see that example. For gender, we had only two categories female and male does we used a single dummy variable 0,1 variable for this. When there are more than two categories the

number of dummy variables that should be used = the number of categories -1. So, the number of dummy variable is number of levels -1.

(Refer Slide Time: 25:33)

Example: Salary vs. Job Grade

- In this example, the categorical variable job grade has 3 levels, 1 (lowest grade), 2, and 3 (highest job grade)

Employee	Job Grade	Salary (\$000)
1	1	7.5
2	3	8.6
3	2	9.1
4	3	10.3
5	3	13
6	1	6.2
7	2	8.7
8	2	9.4
9	3	9.8

You see that there is one example where the job grade is there are three levels. 1, 2, 3, in this example, the categorical variable job grade as three level so, 1, 2, 3, 1 means lowest Grade, 2 means medium and 3 means highest grade. We are going to have three levels in our categorical data three levels are level 1, level 2, level 3.

(Refer Slide Time: 25:54)

Representing 3-level Job Grade using dummy variables

Job_1 and Job_2

Employee's Job Grade	Job Grade	Dummy Variables	
		Job_1	Job_2
1	1	0	
2	0	1	
3	0	0	

Job Grade 3 is the reference category

There are 3 levels and we are going to have only 2 Dummy variables. Job 1 say taken as 1, 0 job2 taken as 0,1 job3 is 0,0. So now, we can say this 0, 0 is taken as a reference, ok. So, the

presence of 1,0 will explain category 1; 0,1 will explain category 2; 0,0 will explain category 3. So here what is happening is there are 3 levels. But we are going to have only two dummy variable dummy variables. Dummy variable 1 and dummy variable 2.

(Refer Slide Time: 26:33)

Employee	Grade	Salary	Job_1	Job_2
1	1	7.5	1	0
2	3	8.6	0	0
3	2	9.1	0	1
4	3	10.3	0	0
5	3	13	0	0
6	1	6.2	1	0
7	2	8.7	0	1
8	2	9.4	0	1
9	3	9.8	0	0

Now, this is a new data set how this data set can be used for doing dummy variable regression. The interpretation is already I have explained to you now will go for demo of this code which I have shown in our, this presentation.

(Refer Slide Time: 26:49)

```

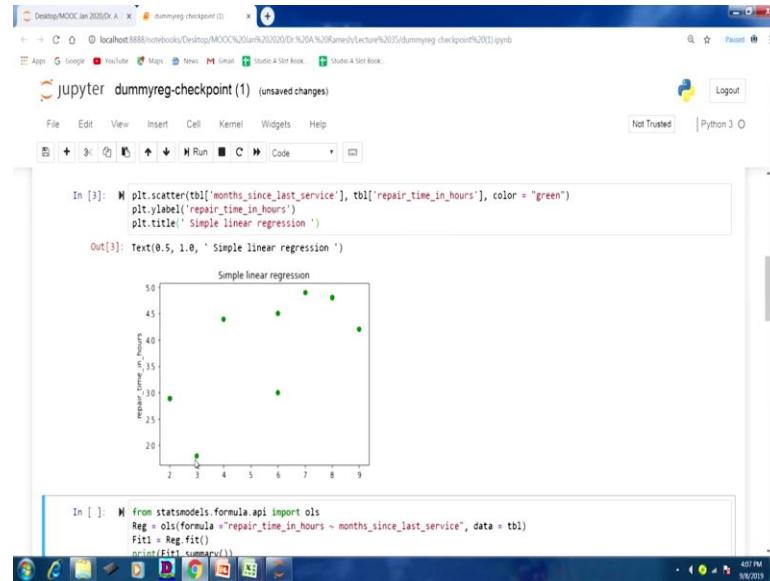
In [1]: import pandas as pd
        import matplotlib as plt
        import statsmodels.formula.api as sm
        from sklearn.linear_model import LinearRegression
        from scipy import stats
        import seaborn as sns
        import numpy as np
        import matplotlib.pyplot as plt
        import statsmodels.api as s
        ...

In [2]: tbl = pd.read_excel('dummy.xlsx')
Out[2]:
servicecall months_since_last_service type_of_repair repair_time_in_hours
0 1 2 electrical 29
1 2 6 mechanical 30
2 3 8 electrical 48
3 4 3 mechanical 18
4 5 2 electrical 29
5 6 7 electrical 49
    
```

I have prepared already code for that person. First I am going to remove this output by clicking kernel restart and clear output. I have cleared the output now I am going to run this one. So as

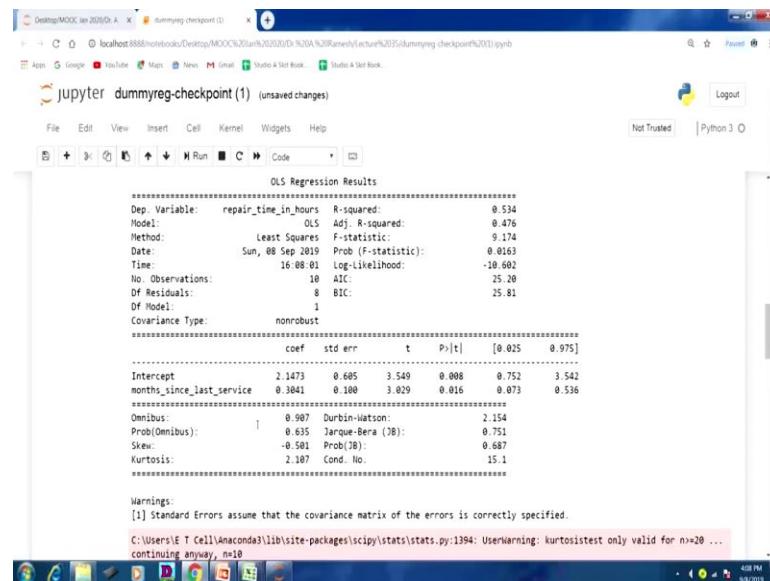
you know, this is Shift Enter so again shift enter this is the data. This data shows service call months since last service type of repair. Next we will go for scatter plot.

(Refer Slide Time: 27:17)



Scatter plot shows that there is a correlation between month since last service and repair time in hours next will go for simple linear regression where were taken only one independent variable.

(Refer Slide Time: 27:35)



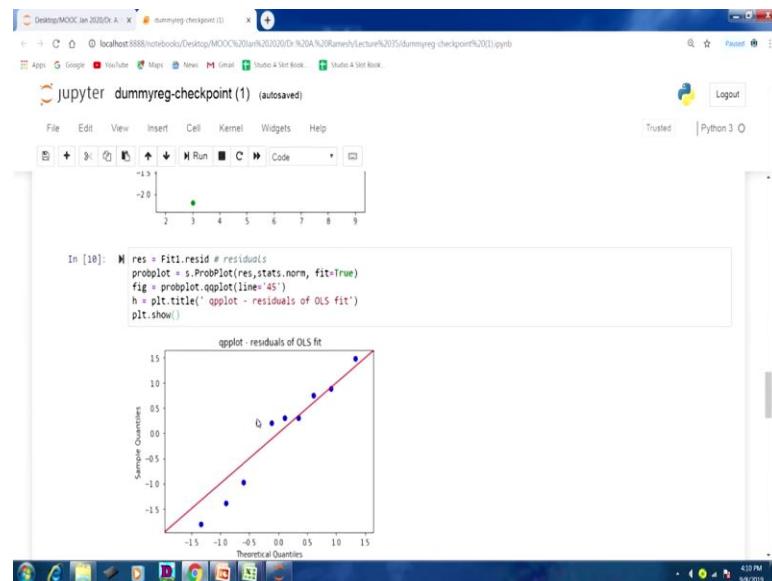
When you look at this here, this is equal to $2.14 + 0.3041 \times 1$. Suppose these variables x_1 . Look at the p-value this p-value is less than 0.05. So this variable is significant value variable, R square also it is good above than 0.5. See, when there is f statistic this is also less than 0.05. So, as a whole model it is valid.

(Refer Slide Time: 28:04)

The screenshot shows a Jupyter Notebook interface. In cell [8], the code `influence = Fit1.get_influence()` and `resid_student = influence.resid_studentized_external` is run. In cell [9], the code `plt.figure()` and `plt.scatter(tbl['months_since_last_service'], resid_student, color = "green")` is run, resulting in a scatter plot where points are green dots. In cell [10], the code `res = Fit1.resid # residuals` and `probplot = s.ProbPlot(res.stats.norm, fit=True)` is run, followed by `fig = probplot.qqplot(line='45')` and `plt.show()`. The plot shows standardized residuals against theoretical quantiles, with a red diagonal line representing the OLS fit.

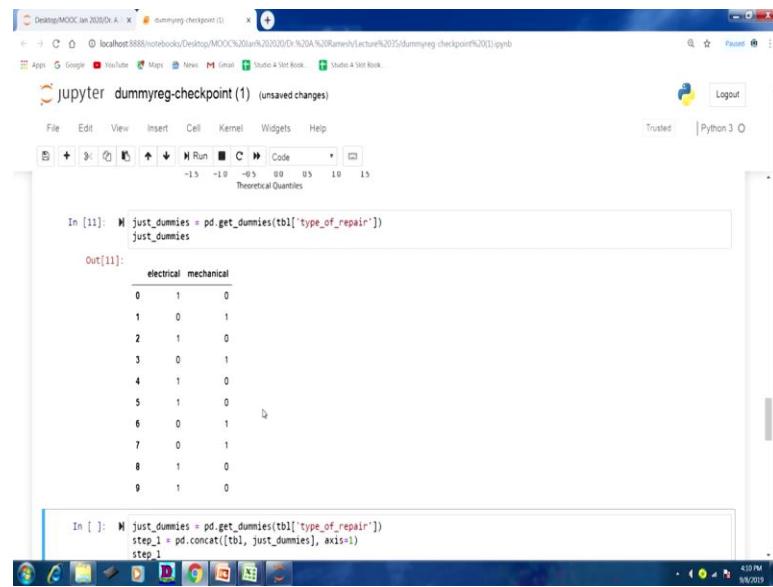
Now, we will plot standardized residual plot. When you look at the standardized residual plot this is the pattern. See there are that some points which are going above -2 how to interpret the standardized residual plot all the points should be between – 2 to + 2. But it seems that there are some variable which goes beyond -2. So it is violating our model assumptions. Now we will go for this q plot.

(Refer Slide Time: 28:38)



These also see that these into some pattern continuously three lines are below this line. There are so many points are above this line. There are also problems in variance of the error variable also.

(Refer Slide Time: 28:51)



The screenshot shows a Jupyter Notebook window titled "jupyter dummyreg-checkpoint (1) (unsaved changes)". The notebook is running on Python 3. In the code cell (In [11]), the following command is run:

```
In [11]: just_dummies = pd.get_dummies(tbl['type_of_repair'])  
just_dummies
```

The output (Out[11]) displays a DataFrame with two columns: "electrical" and "mechanical". The data is as follows:

	electrical	mechanical
0	1	0
1	0	1
2	1	0
3	0	1
4	1	0
5	1	0
6	0	1
7	0	1
8	1	0
9	1	0

In the next cell (In []), the following command is partially visible:

```
In [ ]: just_dummies = pd.get_dummies(tbl['type_of_repair'])  
step_1 = pd.concat([tbl, just_dummies], axis=1)  
step_1
```

Now we will convert the data into dummy variable. This is dummy variable electrical is taken as one mechanical is taken as 0, now after converting into dummy variable to drop the column dummy variable belongs to mechanical. After Dropping we can see this output, for duplicate this now. There is no mechanical column only electrical column is there.

I will do for this data set will go for regression analysis two independent variable one is months underscore since last service another one is type of repair that is electrical is taken as reference. When we look at the p-value the p-value are both independent variables less than 0.05, so the significant model in this equation when you substitute x_2 equal to 1 will get a regression equation for problem related to electrical. When you substitute x_2 equal to 0 will get a regression equation for problem related to mechanical repairing system.

Now we will be going for another problem. This is our second problem, where the salary is the dependent variable. Experience is independent variable gender also independent variable. When we plot that between experience and salary there is a positive relationship. Now will take salary and experience, experience is an independent variable. You see that experiences a significant because less than 0.05. R square is 0.26.

There is no problem in this. Now, we look at the standardized residual plot that most of the points there is not equally plotted, most of the point above zero there is no randomness in the distribution. There seems to be some pattern in the residuals. We will go for checking the normality of the variance error term. See that it is this also following some kind of a pattern and then also not sitting on the exactly the diagonal line.

Now will go for create a dummy variable for the gender. There is one is for female another one is male now will drop this one. So, female is taken as 1 male is taken as zero when you do there is regression analysis where Gender is taken as a female now, you see that Y equal to 9.7 plus (-1.175) female. So, females are getting less salary than the male but look at the P value, when you consider only the gender, the P value is more than 0.05. So this gender variable is not significant.

When you, when you bring another variable is an experience when you look at our previous code, it is only gender is taken gender also it is not significant because the p-value 0.389. Now will take Gender and experience together, let us see what is happening. When you take Gender and experience together, you see that the P value for female is less than 0.05. The experience also, listen 0.05. Both the variables are significant, but the female is getting less salary, when compared to male even though they have equal experience.

Now, what will happen when you reverse the code? Suppose, we have taken female equal to 1 male equal to zero now, what will happen? When you reverse that code send male equal to 1 and female equal to zero what will happen if there will not be any change in the result. Only the sign of usually the male is taken that was - 1.17. Now female is taken as reference. So we are getting only the positive value of 1.17.

Only the difference in the Y intercept, otherwise all interpretations are same. In this lecture by using dummy variable regression I have taken 2 problems with the help of python code I have explained how to do a dummy variable regression and I have also interpreted the result. We know what is the dummy variable regression is sometime the gender is one example for dummy variable regression because there are two possibilities, male and female.

Similarly the job category, Category 1, category 2, category 3, these are dummy variable. For this purpose we have learnt how to do a regression analysis, the next class very important topic that is logistic regression, we are going to see that one before seeing Logistic regression. There is a one principle called maximum likelihood principle. I will explain what is the maximum likelihood principle? With the help of some examples, then we will go per Logistic regression in the next class. Thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture - 36
Maximum Likelihood Estimation - I

In this lecture, we will go to new way of estimating the population parameter. That method is called maximum likelihood estimation. In our previous class, we have estimated the population parameter with the help of least square or we can say with the help of method of moments. This method of estimating population parameter has lot of advantages over that two methods. That we will see in this class.

(Refer Slide Time: 00:50)

Agenda

- This lecture will provide intuition behind the MLE using Theory and examples.

$$\begin{array}{ccc} \bar{x} - \mu & \xrightarrow{\text{LSE}} & \rightarrow \text{MLE} \\ s^2 \rightarrow \sigma^2 \end{array}$$

The agenda for this class is to provide an intuition behind maximum likelihood principle and theory and examples. So what we are going to do, we remember in the previous class with the help of \bar{x} , we have predicted the mean with the help of sample variance, we have predicted the population variance with the help of moment. In the regression model, we have used least square estimate. What you have done in this?

The sum of the square of the error is minimized when we draw the best regression equation. Instead of that one, we are going to use another way of estimating population parameter with the help of maximum likelihood estimation. This is very simple. With the help of this maximum

likelihood estimate, you can estimate parameter of any population, it may be any distribution. It may be binomial; it may be a Poisson. It may be an exponential.

What is the assumption? We are having in the least square estimate is that error term should follow normal distribution. Whenever the error term not following normal distribution, the maximum likelihood estimate is the best way. That we will see in this class.

(Refer Slide Time: 02:02)

Maximum Likelihood Estimation

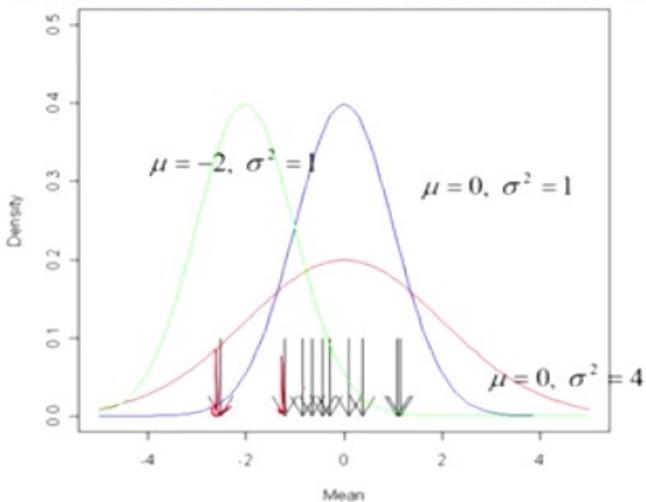
- The method of maximum likelihood was first introduced by R. A. Fisher, a geneticist and statistician, in the 1920s.
- Most statisticians recommend this method, at least when the sample size is large, since the resulting estimators have certain desirable efficiency properties
- Maximum likelihood estimation(MLE) is a method to find most likely density function, that would have generated data.
- MLE requires one to make distribution assumption first.

What is maximum likelihood estimation? The method of maximum likelihood was first introduced by R. A. Fisher, a geneticist and statistician in 1920s. Most statistician recommend this method at least when the sample size is large, since the resulting estimator have certain desirable efficiency properties. Maximum likelihood estimation is a method to find most likely density function that would have generated the data.

So what we can do with the help of this MLE is that which distribution has generated the data that we can find out. Otherwise, this data set is suitable for what kind of distribution? But one assumption we have to have this maximum likelihood estimation is that it requires that one to make distribution assumption first. So in advance, we have to assume which distribution has generated that set of data.

(Refer Slide Time: 03:01)

An intuitive view on likelihood



Let us see the intuitive view on likelihood. See there are some data set there, in the bottom. See there are this data set. We want to know from which normal distribution this data set might have come. There are three possibilities; one is the green line, whose mean is minus 2, variance is 1. The another one is blue, whose mean is 0 and the variance is 1. The last one is mean equal to 0, the variance is 4.

So the most suitable for this one is the blue one, because that covers all the data set. So the purpose of maximum likelihood principle is; suppose there are some data. This data has come from which distribution. So that kind of testing can be done with the help of this. Otherwise, this data set is suitable for what kind of distribution; the other way also. So this is most useful for estimating many population parameters.

(Refer Slide Time: 04:01)

Maximum Likelihood Estimation: Problem

- A sample of ten new bike helmets manufactured by a certain company is obtained. Upon testing, it is found that the **first, third, and tenth** helmets are flawed, whereas the others are not.
- Let $p = P(\text{flawed helmet})$, i.e., p is the proportion of all such helmets that are flawed.
- Define (Bernoulli) random variables X_1, X_2, \dots, X_{10} by

$$X_1 = \begin{cases} 1 & \text{if 1st helmet is flawed} \\ 0 & \text{if 1st helmet isn't flawed} \end{cases} \dots \quad X_{10} = \begin{cases} 1 & \text{if 10th helmet is flawed} \\ 0 & \text{if 10th helmet isn't flawed} \end{cases}$$

Source: Probability and Statistics for Engineering and the Sciences, Jay L Devore, 8th Ed, Cengage

We will take one simple example. With the help of this example, I will explain what is the application of this maximum likelihood estimation? This problem is taken from this book probability and statistics for engineering and sciences by professor Jay L. Devore 8th edition. It is Cengage publications. The problem says a sample of 10 new bike helmets manufactured by a certain company is obtained.

Upon testing, it is found that the first, third, and 10th helmets are flawed; whereas the others are not. Let p is the probability of flawed helmet that is p is the proportion of all such helmets that are flawed. Define Bernoulli random variable X_1, X_2, \dots, X_{10} by; we are going to use $X_1 = 1$, if the helmet is flawed; if there is a defect. $X_1 = 0$, if the helmet is not defective. Like that, if X_{10} value = 1, if the 10th helmet is flawed, 0 if the 10th helmet is not flawed, defective.

(Refer Slide Time: 05:18)

Maximum Likelihood Estimation: Problem

- Then for the obtained sample, $X_1 = X_3 = X_{10} = 1$ and the other seven X_i 's are all zero
- The probability mass function of any particular X_i is $p^{x_i}(1-p)^{1-x_i}$, which becomes p if $x_i = 1$ and $1-p$ when $x_i = 0$
- Now suppose that the conditions of various helmets are independent of one another
- This implies that the X_i 's are independent, so their joint probability mass function is the product of the individual pmf's.

Then, for the obtained sample, say $X_1 = X_3 = X_{10} = 1$, because they are already given; only the first and third and 10th helmet have some defect and the other seven X_i 's are all 0. The values are 0. The probability mass function of any particular X_i is $p^{x_i}(1-p)^{1-x_i}$, which becomes p if X_i equal to 1 and $1-p$ when X_i equal to 0. Now, suppose the conditions of various helmets are independent of one another, because this assumption is very important.

If there is independent, we can find out their joint distribution. This implies that X_i 's are independent, so that their joint probability mass function is the product of their individual probability mass function.

(Refer Slide Time: 06:19)

Maximum Likelihood Estimation: Binomial Distribution

- Joint pmf evaluated at the observed X_i 's is
$$f(x_1, \dots, x_{10}; p) = p(1-p)p \cdots p = p^3(1-p)^7 - (1)$$
- Suppose that $p = .25$. Then the probability of observing the sample that we actually obtained is $(.25)^3(.75)^7 = .002086$.
- If instead $p = .50$, then this probability is $(.50)^3(.50)^7 = .000977$.
- For what value of p is the obtained sample most likely to have occurred?
- That is, for what value of p is the joint pmf (eq 1) as large as it can be?
- What value of p maximizes (eq 1)

Since it is joint probability mass function, see that we have multiplied for all possibilities, p_i into $(1 - p_i)$ by considering all possibilities. So when you simplify that p^3 into $(1 - p)^7$. This equation is 1. Suppose, in that equation, this left hand side, this value is called maximum. This is a likelihood value. Whatever is in the left hand side, I will define it later, what is the likelihood value. The left hand value is called likelihood value.

Suppose with the $p = 0.25$, then the probability of observing the sample that we actually obtained is 0.002086. So like that, we can supply different p values. Suppose, instead of 0.25, you supply $p = 0.5$, then the probability is 0.0097. You see that, when it is a 0.25, 0.002, when it is a 0.5, it has become very low. So in between this 0.25 and 0.50, we are going to get the value of p that will maximize our left hand side value.

For what value of p is obtained sample, most likely to have occurred? That was the question. What is that? For what value of p is obtained sample most likely to have occurred; that is for what value of p is the joint pmf; this one, as large as it can be; otherwise what value of p maximizes equation 1. That p value is nothing but your likelihood value.

(Refer Slide Time: 08:07)

Maximum Likelihood Estimation: Binomial Distribution

- Figure shows a graph of the likelihood (eq 1) as a function of p .
- It appears that the graph reaches its peak above $p = .3$ = the proportion of flawed helmets in the sample.

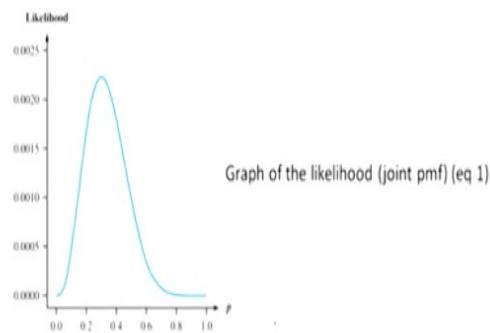


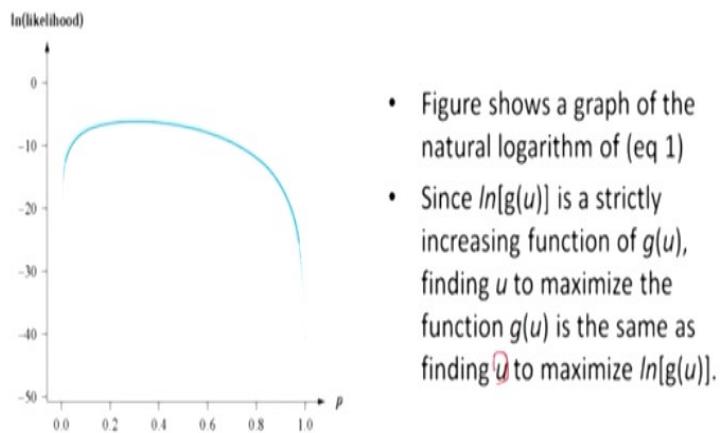
Figure shows a graph of likelihood of that value of equation 1 as a function of p . So what happened in x axis, you have taken different value of p . Previously, just for one case, we have taken p value 0.25 and 0.50. See when it is 0.25, this was the likelihood value, 0.5, this was the

likelihood value. So we are good to supply, draw a graph by supplying different value of p in equation 1 that we have to find out the likelihood value.

This figure shows a graph of likelihood as a function of p. It appears that the graph reaches its peak above 0.3, when the value is 0.3; the graph reaches its peak, equal to the proportion of flawed helmets in the sample. Now what you are going to do? We are going to take log of this function.

(Refer Slide Time: 09:04)

Graph of the natural logarithm of the likelihood



I will explain; there was a reason for that. Graph of the natural logarithm of likelihood. Figure shows graph of the natural logarithm of equation 1. Since the logarithm of $g[u]$ is strictly increasing function of $g[u]$, finding u to maximize the function g of u is the same as finding u to maximize $\log g$ of u . So what is happening is, whether g of u and logarithm of g of u is the same. This figure shows a graph of the natural logarithm of equation 1.

Since $\log g[u]$ is strictly increasing function of g of u , finding u to maximize the function of g of u is the same as finding u to maximize $\log g$ of u . So the u is same, whether it is g of u or $\log g$ of u .

(Refer Slide Time: 10:04)

Maximum Likelihood Estimation: Binomial Distribution

- We can verify our visual impression by using calculus to find the value of p that maximizes (eq 1).
- Working with the natural log of the joint pmf is often easier than working with the joint pmf itself, since the joint pmf is typically a product so its logarithm will be a sum.
- Here $\ln[f(x_1, \dots, x_{10}; p)] = \ln[p^3(1-p)^7]$
- $= 3\ln(p) + 7\ln(1-p)$

$$\begin{aligned} & \ln p^3 + \ln(1-p)^7 \\ & 3\ln(p) + 7\ln(1-p) \end{aligned}$$

We can verify our visual impression by using calculus to find out the value of p that maximizes equation 1. Working with natural logarithm of the joint probability mass function is often easier than working with the joint pmf itself. Since the joint pmf is typically a product, so the logarithm will be a sum. That is the advantage of taking log of that. So what will happen previously in equation 1, we got pq multiplied by $1 - p$ power 7. I am going back here; this one.

We are going to take log of this. When you take log of this, it will become, because there is a multiplication, so this will become $\log(p^3) + \log((1-p)^7)$. So this will become $3\log(p) + 7\log(1-p)$.

(Refer Slide Time: 11:12)

Maximum Likelihood Estimation: Binomial Distribution

$$\text{Thus } \frac{d}{dp} \{\ln[f(x_1, \dots, x_{10}; p)]\} = \frac{d}{dp} \{3\ln(p) + 7\ln(1-p)\}$$

$$= \frac{3}{p} + \frac{7}{1-p}(-1)$$

$$\frac{dy}{dx} = 0$$

$$= \frac{3}{p} - \frac{7}{1-p}$$

$$\frac{dy}{dx} = 0$$

$$\frac{d}{dp} = 0$$

Next one is that functions, we have to see when the value become maximum? We know that in our school, we might have studied; you see to find out the maximum value, maxima-minima. For example, the maximum value, if you say dy/dx equal to 0; then (d^2y/dx^2) will be less than 0 means that point will become the maximum. So this equation, this is the function of p. So we are going to differentiate that log function with respect to p.

So when you differentiate this one, so $3 \log p = (1/p)$, so $(3/p) + 7$ is a constant. Log of $1-p$ is, this is differentiation, log of x equal to 1 by x. So 1 divided by $(1-p)$, again you have to differentiate this function. Differentiation of differentiation, so $0 - 1$, so it will be -1 . So $(3/p) - 7$ divided by $(1-p)$. So this equations, we have to equate it to 0, because we know (d/dp) should be equal to 0. Then, we have to find out the p. So that value, the function will get maximized.

(Refer Slide Time: 12:32)

Interpretation

- Equating this derivative to 0 and solving for p gives
 $3(1-p) = 7p$, from which $3 = 10p$ and so $p = 3/10 = .30$ as conjectured
- That is, our point estimate is $p = .30$.
- It is called the *maximum likelihood estimate* because it is the parameter value that maximizes the likelihood (joint pmf) of the observed sample
- In general, the second derivative should be examined to make sure a maximum has been obtained, but here this is obvious from Figure

Equating these derivatives to 0 and solving for p, it gives 3 into $1-p$ equal to $7p$. So 3 equal to $7p$, so $p = 0.3$ is conjectured. So now what is happening? Previously, we have substituted different values. Now we are using the concept of maxima, we have realized that when the $p = 0.3$, the function gets maximized. So it is called the maximum likelihood estimate, because it is the parameter value that maximizes the likelihood of the observed sample.

So this $p = 0.3$ will be nothing but the; this is an estimate for the population. In general, second derivative should be maximum to make sure the maximum has been obtained, but here this is

obvious from the figure. So actually we have to differentiate one more time and we have to see whether it has become negative or not, because by looking at the figure, it seems that that point is maximum. So what is happening; this value $p = 0.3$ is called the maximum likelihood estimate.

So what is happening, the binomial distribution of the population parameter p , we have estimated it is 0.3. So the advantage of this maximum likelihood function is, it is helping to estimate parameter of any distribution.

(Refer Slide Time: 14:01)

Maximum Likelihood Estimation: Binomial Distribution

- Suppose that rather than being told the condition of every helmet, we had only been informed that three of the ten were flawed.
- Then we would have the observed value of a binomial random variable $X =$ the number of flawed helmets.
- The pmf of X is $\binom{10}{x} p^x (1-p)^{10-x}$. For $x = 3$, this becomes $\binom{10}{3} p^3 (1-p)^7$.
- The binomial coefficient $\binom{10}{3}$ is irrelevant to the maximization, so again $p = 0.30$.

Suppose, that rather than being told the condition of every helmet, we had only been informed that three of the 10 were flawed. Then, we would have to observe the value of binomial random variable X equal to the number of flawed helmets when you substitute $10, X; {}^{10}C_x p^x$ into $(1-p)^{10-x}$. When you substitute $x = 3$, this is ${}^{10}C_3 p^3$ into $(1-p)^7$. We do not bother about the coefficient ${}^{10}C_3$, because that is not a function; that is just a constant. So what they say, the binomial coefficient ${}^{10}C_3$ is irrelevant to maximization. So again, the $p = 0.3$.

(Refer Slide Time: 14:44)

Maximum Likelihood Function Definition

- Let X_1, X_2, \dots, X_n have joint pmf or pdf
$$f(x_1, x_2, \dots, x_n; \theta_1, \dots, \theta_m) \quad (a)$$

- Where the parameters $\theta_1, \dots, \theta_m$ have unknown values. When x_1, \dots, x_n are the observed sample values and (a) is regarded as a function of $\theta_1, \dots, \theta_m$, it is called the **likelihood function**.

- The maximum likelihood estimates (mle's) $\hat{\theta}_1, \dots, \hat{\theta}_m$ are those values of the θ 's that maximize the likelihood function, so that

$$f(x_1, x_2, \dots, x_n; \hat{\theta}_1, \dots, \hat{\theta}_m) \geq f(x_1, x_2, \dots, x_n; \theta_1, \dots, \theta_m) \text{ for all } \theta_1, \dots, \theta_m$$

- When the \hat{X}_i 's are substituted in place of the x_i 's, the **maximum likelihood estimators result**.

Next, we will define, what is maximum likelihood function. There are two terms there; one is likelihood function, next one is maximum likelihood function. First I will say what is likelihood function, then we will go to what is maximum likelihood function. Let X_1, X_2, \dots, X_n have a joint probability mass function or probability density function; call it as f of $x_1, x_2, \dots, x_n ; \Theta_1, \Theta_2, \dots, \Theta_m$, where the parameters $\Theta_1, \Theta_2, \dots, \Theta_m$ have unknown values.

Here the parameter is $\Theta_1, \Theta_2, \dots, \Theta_m$, unknown values where x_1, x_2, \dots, x_n are the observed sample values, then this equation a is regarded as the function of $\Theta_1, \Theta_2, \dots, \Theta_m$, it is called likelihood function. So this is a likelihood function. So this function is likelihood function. The maximum likelihood estimates $\hat{\Theta}_1, \hat{\Theta}_2, \dots, \hat{\Theta}_m$ are those values of Θ 's that maximizes the likelihood function.

So that $f(x_1, x_2, \dots, x_n ; \hat{\Theta}_1, \hat{\Theta}_2, \dots, \hat{\Theta}_m)$ is greater than or equal to $f(x_1, x_2, \dots, x_n ; \Theta_1, \Theta_2, \dots, \Theta_m)$ for all values of $\Theta_1, \Theta_2, \dots, \Theta_m$. When the \hat{X}_i 's are substituted in place of x_i 's, the maximum likelihood estimates result. So what you have to do with that one? In the X_i 's we have to substitute x_i 's that will be the maximum likelihood estimate result. So what we are doing here?

We are finding joint probability mass function, then with the help of sample values, we are predicting the population parameter.

(Refer Slide Time: 16:48)

Interpretation

- The likelihood function tells us how likely the observed sample is as a function of the possible parameter values.
- Maximizing the likelihood gives the parameter values for which the observed sample is most likely to have been generated—that is, the parameter values that “agree most closely” with the observed data.

How will you interpret that one? The likelihood function tells us how likely the observed sample is as a function of possible parameter values. So maximizing the likelihood gives the parameter values, for which the observed value is most likely to have been generated. That is, the parameter values that agree most closely with the observed data. Otherwise, we can say in other way, that this data set is more suitable for what kind of distributions or what kind of models.

(Refer Slide Time: 17:22)

Estimation of Poisson Parameter

- Suppose we have data generated from a Poisson distribution. We want to estimate the parameter of the distribution
- The probability of observing a particular random variable is $P(X; \mu) = \frac{e^{-\mu} \mu^X}{X!}$
- Joint likelihood by multiplying the individual probabilities together

$$P(X_1, X_2, \dots, X_n; \mu) = \frac{e^{-\mu} \mu^{X_1}}{X_1!} \times \frac{e^{-\mu} \mu^{X_2}}{X_2!} \times \dots \times \frac{e^{-\mu} \mu^{X_n}}{X_n!}$$

$$L(\mu; \mathbf{X}) = \prod_i e^{-\mu} \mu^{X_i}$$

$$L(\mu; \mathbf{X}) = e^{-n\mu} \mu^{n\bar{X}}$$

Now we will go for estimation of Poisson parameter. Suppose, we have data generated from a Poisson distribution, we want to estimate the parameter of Poisson distribution. The Poisson distribution is having only one parameter, because in Poisson distribution, it is an unique

parametric distribution; it has only one parameter, that is where the mean and variance is same. The probability of observing a particular random variable $P(X; \mu) = (e^{-\mu} \mu^X)/X!$

So joint likelihood by multiplying the individual probabilities together, so what we will do the first step is we have to find out the joint probability function. So $(e^{-\mu} \mu^{X_1})/X_1!$ multiplied by $(e^{-\mu} \mu^{X_2})/X_2!$ and so on multiply $(e^{-\mu} \mu^{X_n})/X_n!$. So this can be written as product of $(e^{-\mu} \mu^{X_i})$ because it is a product, there is an end time. When you expand it, so $e^{-n\mu}$, because it will become up to n times, so $\mu^{n\bar{X}}$. Next, we have to take the log of this, we will see that.

(Refer Slide Time: 18:45)

Estimation of Poisson Parameter

- Note in the likelihood function the factorials have disappeared.
- This is because they provide a constant that does not influence the relative likelihood of different values of the parameter
- It is usual to work with the **log likelihood** rather than the likelihood.
- Note that maximising the log likelihood is equivalent to maximising the likelihood.

$$\begin{aligned} L(\mu; \mathbf{X}) &= e^{-n\mu} \mu^{n\bar{X}} && \text{Take the natural log of the likelihood function} \\ \ell(\mu; \mathbf{X}) &= -n\mu + n\bar{X} \log \mu && \text{Find where the derivative of the log likelihood is zero} \\ \frac{d\ell}{d\mu} &= -n + \frac{n\bar{X}}{\mu} && \text{Note that here the MLE is the same as the moment estimator} \\ \hat{\mu} &= \bar{X} \end{aligned}$$

Note that, the likelihood function that factorials have disappeared. We will not bother about the factorials, because that is not going to affect the result. This is because they provide a constant that does not influence the relative likelihood of different values of the parameter whether we use the constant or not, that is not required, because that will not affect our end result. It is usual to work with log likelihood rather than likelihood, because we have seen previously.

When you take log of likelihood, the differentiation is easy. Note that, maximizing the log likelihood is equivalent to the maximizing likelihood. This also, we have seen in the previous slide. So this was the likelihood function. You take log of that one. So e power, when you take log of e to the power $-n\mu$ is $-\mu$, because it is the product, in log it will become sum, sum $n\bar{X}$ bar log of μ . Now you differentiate with respect to μ .

When you differentiate it and equate it to 0, then you are getting \bar{X} equal to μ . So what is the result is, the sample mean is the best estimate to predict the population mean of a Poisson distribution.

(Refer Slide Time: 20:04)

Estimation of exponential distribution Parameter

- Suppose X_1, X_2, \dots, X_n is a random sample from an exponential distribution with parameter λ . Because of independence, the likelihood function is a product of the individual pdf's:

$$\begin{aligned} f(x_1, \dots, x_n; \lambda) &= (\lambda e^{-\lambda x_1}) \cdot \dots \cdot (\lambda e^{-\lambda x_n}) \\ &= \lambda^n e^{-\lambda \sum x_i} \end{aligned}$$

- The natural logarithm of the likelihood function is

$$\ln[f(x_1, \dots, x_n; \lambda)] = n \ln(\lambda) - \lambda \sum x_i$$

Now we will go for another distribution that is estimation of exponential distribution parameter. Suppose, X_1, X_2, X_n is a random sample from an exponential distribution with the parameter lambda. Because of independence, the likelihood function is the product of individual pdf's. Here also $\lambda e^{-\lambda x_1}$ will extend it, $\lambda e^{-\lambda x_2}$ up to, you have to multiply $\lambda e^{-\lambda x_n}$.

So when you simplify that, it will become $\lambda^n e^{-\lambda \sum x_i}$. When you take log of this, it will become $n \ln(\lambda) - \lambda \sum x_i$. Then this has to be equated to 0.

(Refer Slide Time: 20:59)

Estimation of exponential distribution Parameter

- Equating $(d/d\lambda)[\ln(\text{likelihood})]$ to zero results in $n/\lambda - \sum x_i = 0$, or $\lambda = n/\sum x_i = 1/\bar{x}$.
- Thus the MLE is $\hat{\lambda} = 1/\bar{x}$;

So when you equate it to 0, so lambda becomes lambda equal to n divided by sigma Xi that is nothing but the inverse of the sample mean. So this was the result. So what is happening is, now the inverse of sample mean is nothing but the mean of our exponential distribution.

(Refer Slide Time: 21:24)

Estimation of parameters of Normal Distribution

- Let X_1, \dots, X_n be a random sample from a normal distribution.
- The likelihood function is

$$f(x_1, \dots, x_n; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/(2\sigma^2)} \cdot \dots \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/(2\sigma^2)}$$
$$= \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} e^{-\sum(x_i-\mu)^2/(2\sigma^2)}$$

- so

$$\ln[f(x_1, \dots, x_n; \mu, \sigma^2)] = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum(x_i - \mu)^2 \quad \text{--- } \textcolor{red}{1}$$

Now we will go for estimation of parameter of a normal distribution. This was very interesting because we can say normal distribution is the father of all the distributions. Many time, if you are not knowing the nature of the distribution, you can assume that it follows normal distribution. As usual, the likelihood function for a normal distribution is, we know that the pdf, probability density function is $(1/\sqrt{2\pi\sigma^2}) e^{-(x_i-\mu)^2/(2\sigma^2)}$.

So like that, this is term 1, term 2, up to nth term we can go for that. So term 1, that is when it is x_1 , when you substitute x_2, x_3 , so you will get different n terms. So that is probability mass function. Joint probability function, so when you simplify that it is $(1 \text{ divided by } 2\pi\sigma^2)$ to the power $n/2$, into e to the power $(-\sum(x_i - \mu)^2 \text{ divided by } 2\sigma^2)$. What will happen? When you take log of this, this is $(n/2)(\ln(1) - \ln(2\pi\sigma^2))$.

So what will happen, log of 1 minus, because log 1 is 0, so it will become $0 - \ln(2\pi\sigma^2)$, because it is x power n . So $-n$ by $2n \log 2\pi\sigma^2 - e$ to the power, this one will come in that value itself, because $2\sigma^2$, sigma of $x_i - \mu$ whole square. Now what has to be done? This is the log value of likelihood function. This function, this equation has to be partially derivated with respect to μ and σ^2 and equate it to 0.

(Refer Slide Time: 23:19)

Estimation of parameters of normal distribution

- To find the maximizing values of μ and σ^2 , we must take the partial derivatives of $\ln(f)$ with respect to μ and σ^2 , equate them to zero, and solve the resulting two equations.
- Omitting the details, the resulting MLE's are

$$\hat{\mu} = \bar{X} \quad \hat{\sigma}^2 = \frac{\sum(X_i - \bar{X})^2}{n}$$

- The MLE of σ^2 is not the unbiased estimator, so two different principles of estimation (unbiasedness and maximum likelihood) yield two different estimators

Then, you will get the parameter. To find the maximizing value of μ and σ^2 , we must take the partial derivatives of the previous function with respect to μ and σ^2 and equate them to 0 and solve the resulting two equations. There are a lot of details, omitting the details, we will get this result. What does this result says? With the help of sample mean, we can predict the population mean.

With the help of this one, look at this one; this term is the sample variance, we can predict the population variance. So this was the outcome of, you remember, this was the result of our central

limit theorem also. We can prove that central limit theorem by using this maximum likelihood estimate. But one point you should be very careful, the maximum likelihood estimate of sigma square is not the unbiased estimator.

Actually, we should look for unbiased estimator, but here it is not unbiased estimator. So, two different principles of estimation, unbiasedness and maximum likelihood yield two different estimators. In this class, I have started the intuitive meaning of maximum likelihood principle. Then, I have explained how to find out the population parameter of different distributions. First, I have seen how to predict the population parameter of binomial distribution.

Next we have seen how to predict the parameter of Poisson distribution. Then, next we have predicted the population parameter of exponential distribution. At last, we have predicted population parameter of normal distributions. In the next class, we will take one example for predicting the parameter of normal distribution. Thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture - 37
Maximum Likelihood Estimation - II

In the previous class, we have seen estimation of population parameter with the help of maximum likelihood principles. In this class, we will take some; two examples. One is to estimate the population parameter of normal distribution. Second one is to estimate the population parameter of a regression equation.

(Refer Slide Time: 00:48)

Agenda

- Python demo for estimation of population parameters for regression equation

At the end, we will have a demo by using Python. What is the agenda for this class is Python demo of estimation of population parameters for regression equation. Let us take one example.

(Refer Slide Time: 01:00)

Example1: Estimation of parameters of normal distribution

- Let us explain basic idea of MLE using simple problems.
- Let us make assumption that variable x follows normal distributed
- Density function of normal distribution with mean μ and variance σ^2 is given by:

Id	x
1	1
2	4
3	5
4	6
5	9

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \text{ for } -\infty < x < \infty$$

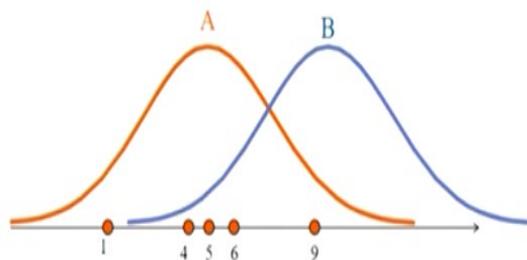
Id is 1, 2, 3, 4. Example 1 is estimation of population parameter of a normal distribution. Let us explain basic idea of maximum likelihood estimation using a simple problem. Let us make assumption that variable X follows normal distribution. The value of variable is 1, 4, 5, 6, 9. The density function of normal distribution with mean μ and variance σ^2 is given by (1 divided by root of $2\pi\sigma^2$) e to the power $-(x_i-\mu)^2$ divided by $2\sigma^2$.

The range of x is between minus infinity to plus infinity. So in this equations, we going substitute these x values, even I am going to multiply that, then we are going to take log of that. Then, we have to partially differentiate with respect to x and mu and equate it to 0, then we will get the population parameter.

(Refer Slide Time: 02:05)

Example 1: Estimation of parameters of normal distribution

- The data is plotted on a horizontal line
- Think which distribution, either A or B, is more likely to have generated the data?



Suppose the data is plotted on a horizontal line, this way. Think which distribution either A or B is more likely to have generated the data. Pause the video, you can think.

(Refer Slide Time: 02:21)

Interpretation

- Answer to this question is A, because the data are clustered around the center of the distribution A, but not around the center of the distribution B
- This example illustrates that, by looking at the data, it is possible to find the distribution that is most likely to have generated the data
- Now, I will explain exactly how to find the distribution in practice

The answer to the question is A, because the data are clustered around the center of the distribution A, but not around the center of the distribution B. This example illustrates that by looking at the data, it is possible to find the distribution that is most likely to have generated the data. Now, I will explain exactly how to find the distribution in practice.

(Refer Slide Time: 02:48)

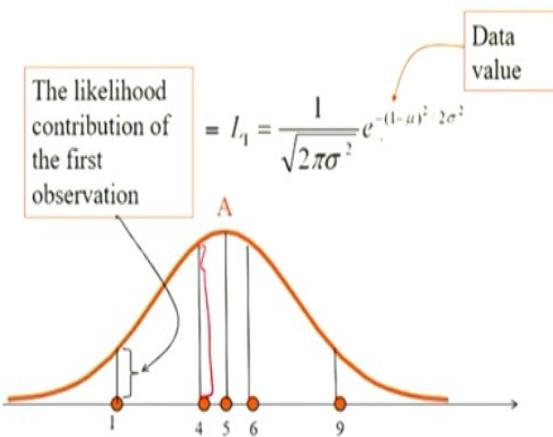
The illustration of the estimation procedure

- MLE starts with computing the likelihood contribution of each observation
- The likelihood contribution is the height of the density function.
- We use L_i to denote the likelihood contribution of i^{th} observation.

Maximum likelihood estimate starts with computing the likelihood contribution of each observation. The likelihood contribution is the height of the density function. I will show you in the next slide. We use L_i to denote the likelihood contribution of i th observation.

(Refer Slide Time: 03:07)

Graphical illustration of likelihood contribution



Look at this picture. The observation 1 has contributed up to this much height. The likelihood contribution of first observation is this much. The likelihood contribution of second observation is this much; similarly, for third, fourth and fifth. For example, this is the first one, is because there is x . Instead of x , we have substituted 1. For second data set, instead of x , we have to substitute 4. For third, 5, the next one 6, the next one 9.

(Refer Slide Time: 03:40)

The illustration of the estimation procedure

- Then, you multiply the likelihood contributions of all the observations. this is called the likelihood function. We use the notation L
- Likelihood function $L = \prod_{i=1}^n L_i$ This notation means you multiply from $i=1$ through n
- In our example, $n=5$

Then, you multiply the likelihood contribution of all the observations. This is called likelihood function. We denote that by L. So likelihood function is the product of L_i . This notation means that you multiply from 1 to n. In our example $n = 5$, so you have to multiply five times.

(Refer Slide Time: 04:03)

The illustration of the estimation procedure

- In our example, the likelihood function looks like:

$$\begin{aligned}
 L(\mu, \sigma) &= \prod_{i=1}^5 L_i = L_1 \times L_2 \times L_3 \times L_4 \times L_5 \\
 &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(1-\mu)^2/\sigma^2} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(4-\mu)^2/\sigma^2} \\
 &\quad \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(5-\mu)^2/\sigma^2} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(6-\mu)^2/\sigma^2} \\
 &\quad \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(9-\mu)^2/\sigma^2}
 \end{aligned}$$

- The likelihood function depends on mean μ and variance σ^2

So that likelihood function, that is defined by with the help of mu and sigma, is the product of $i = 1$ to 5 , L_i , when you expand this one, L_1 multiplied by L_2 multiplied by L_3 and L_4 and L_5 . So it has come from the normal distribution, we have assumed that it has come from the normal distribution. We know that probability density function of normal distribution is 1 divided by root of $2\pi\sigma^2$ e to the power minus.

For first data set it is $1 - \mu$ whole square divided by σ^2 multiplication 1 divided by root of $2\pi\sigma^2$ for second data set e to the power $-4 - \mu$ whole square divided by σ^2 up to all data set. The last data set will be e to the power $-9 - \mu$ whole square divided by σ^2 . So the likelihood function depends on the value of μ and σ^2 , because look at this. So here, the likelihood function is in terms of μ and σ^2 .

(Refer Slide Time: 05:04)

The illustration of the estimation procedure

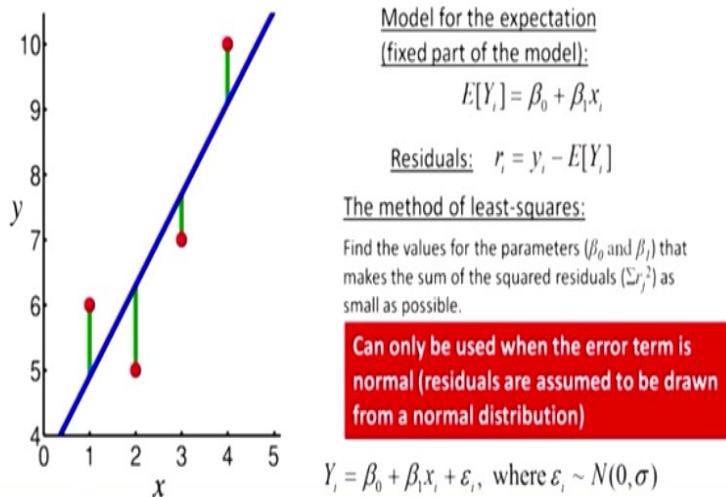
- The value of mean μ and σ that maximise the likelihood function can be found with the help of python.
- The values of mean μ and σ which are obtained this way are called the maximum likelihood estimators of mean μ and σ
- Most of the MLE cannot be solved 'by hand'. Thus, you need to write an iterative procedure to solve it on computer

In the previous slide, we have found the joint probability density function for different value of x , we found the joint probability density function of normal distribution. So what we have to do? We have to take the log of that function, that partially we have to differentiate with respect to μ and σ and equate it to 0, then you will get the population parameter, that is μ and σ . So the value of the mean μ and σ that maximizes the likelihood function can be found with the help of Python.

At the end of the class, I am going to show one example for regression equation. The values of mean μ and σ , which are obtained this way are called maximum likelihood estimator of mean μ and σ . Most of the MLE cannot be solved by hand. That is the maximum likelihood estimate. Thus, you need to write an iterative procedure to solve with computer. That I will take one demo at the end of the class.

(Refer Slide Time: 06:01)

Method of Least-squares vs MLE



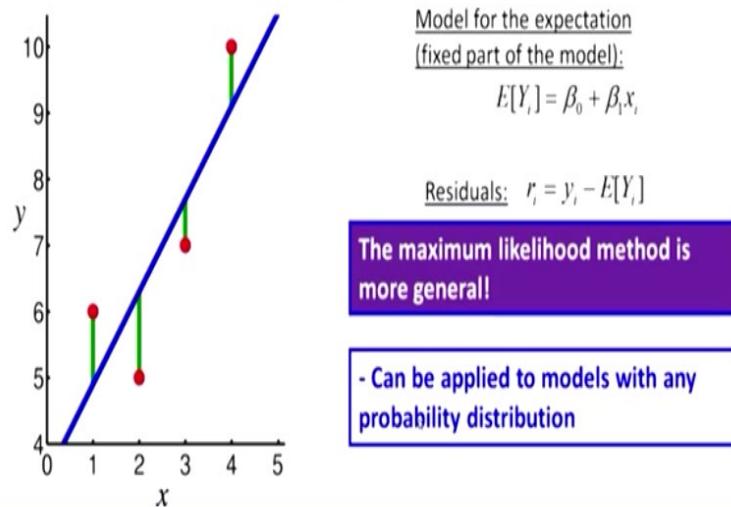
Now we will predict the parameter of a simple regression equation. When I am starting the regression, I have explained the parameter of regression equation was obtained by having this assumption called least square method, where the sum of the square of the error is minimized. So model for the expectation, fixed part of the model, so expected value of Y is beta 0 + beta 1x. The residual is actual minus predicted value y_i – expected value.

The method of least square, what we have done, we have found the values of the parameter beta 0 and beta 1, that makes the sum of the squared residuals as small as possible. This method of least square is applicable only when the error term is normal. That is, residuals are assumed to be drawn from a normal distribution. Whenever this assumption is getting violated, we cannot go for least square method.

We should go for some other method that is maximum likelihood estimate, because we know that the next class, we are going to study about the logistic regression, error term will not follow normal distribution.

(Refer Slide Time: 07:19)

Method of Least-squares vs MLE



So the maximum likelihood estimate can be applied to models with any probability distribution.

That was the advantage of this maximum likelihood method.

(Refer Slide Time: 07:27)

Estimation of Regression Parameter

- We are interested in estimating a model like this:

$$y = \beta_0 + \beta_1 x + u$$

- Estimating such a model can be done using MLE

Now, we will estimate the parameter of a regression equation. We are interested in estimating a model like this, where $y = \beta_0 + \beta_1 x + u$. This u is error term. Estimating such model can be done using maximum likelihood estimation.

(Refer Slide Time: 07:50)

Estimation of Regression Parameters

- Suppose that we have the following data and we are interested in estimating the model:
 $y = \beta_0 + \beta_1 x + u$
- Let us make an assumption that u follows the normal distribution with mean 0 and variance σ^2

Id	Y	X
1	2	1
2	6	4
3	7	5
4	9	6
5	15	9

Suppose, that we have the following data, where x is given independent variable, y is given dependent variable. We are interested in estimating the population parameter β_0 , β_1 . Let us make an assumption that the error term follows normal distribution with mean 0 and variance σ^2 .

(Refer Slide Time: 08:13)

Estimation of Regression Parameters

- We can write the model as :
 $u = y - (\beta_0 + \beta_1 x)$
- This means that $y - (\beta_0 + \beta_1 x)$ follows the normal distribution with mean 0 and variance σ^2
- The likelihood contribution of each data point is the height of the density function at the data points $(y - \beta_0 - \beta_1 x)$

We know that what is error? Error is actual minus predicted value. So the actual value is y minus predicted value is $\beta_0 + \beta_1 x$. This means that $y - \beta_0 - \beta_1 x$ follows the normal distribution with mean 0 and the variance σ^2 . The likelihood contribution of each data point is the height of the density function or the data point where $y = -\beta_0 - \beta_1 x$, because nothing but we brought this minus inside. This was the example.

(Refer Slide Time: 08:49)

Estimation of Regression Parameters

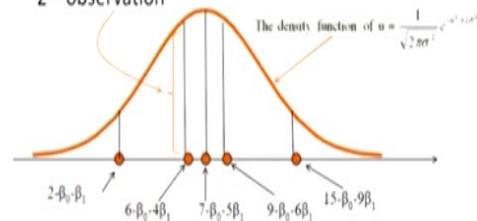
- The likelihood contribution in this example, of the 2nd observation is given

by:

$$L_2 = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(6-\beta_0-4\beta_1)^2/2\sigma^2}$$

Data point

The likelihood contribution of the 2nd observation



When you look at this, for the first data set, it is 2 – beta 0, because 2 when I go for first data set, the y value is 2, x value is 1. So it becomes 2 – beta 0 – beta 1. For second data set, y value is 6, 6 – beta 0 -, x value is 4, 4 beta 1 and so on. So this height is the contribution of each observation on the likelihood function. So likelihood contribution in this example of the second observation is given by, second observation is y value is 6, x values is 4.

So 1 divided by root of 2 pi sigma square e to the power – 6 – beta 0 – 4 beta 1 square divided by 2 sigma square. So the density function u equal to, we can simplify this way. So we will go to next one for other data set.

(Refer Slide Time: 09:46)

Estimation of Regression Parameters

- Then the likelihood function is given by

Id	Y	X
1	2	1
2	6	4
3	7	5
4	9	6
5	15	9

$$\begin{aligned}
 L(\beta_0, \beta_1, \sigma) &= \prod_{i=1}^n L_i = L_1 \times L_2 \times L_3 \times L_4 \times L_5 \\
 &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(2-\beta_0-\beta_1)^2/\sigma^2} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(6-\beta_0-4\beta_1)^2/\sigma^2} \\
 &\quad \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(7-\beta_0-5\beta_1)^2/\sigma^2} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(9-\beta_0-6\beta_1)^2/\sigma^2} \\
 &\quad \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(15-\beta_0-9\beta_1)^2/\sigma^2}
 \end{aligned}$$

- The likelihood function is a function of β_0 , β_1 and σ .

We have done in the previous slide only for these values. Now we will expand that function for all the data set. So the product should go to i to n, L1, L2, L3 up to L5. See for first data set, this is 1 divided by root of 2 pi sigma square . e to the power -2 – beta 0 – beta 1 whole square divided by sigma square. For second data set, the same thing 6 – 4, for third one is 7, 5, for the fourth one is 9, 6, for the fifth one is 15, 9.

So this function is likelihood function. Generally, what we have to do? We have to take the log of this, then we have to partially differentiate with respect to beta 0, beta 1 and sigma and equate it to 0, then you will get the estimation of beta 0, beta 1 and sigma.

(Refer Slide Time: 10:46)

Estimation of Regression Parameters

- You choose the values of β_0 , β_1 and σ that maximizes the likelihood function.

You choose the value of beta 0, beta 1 and sigma that maximizes the likelihood function. So what we are going to do? We are going to, for the same problem, with the help of Python, we are going to predict this beta 0, beta 1 and sigma with the help of data set, which we have considered. We will switch to Python.

(Refer Slide Time: 11:06)

Python Demo for MLE

```
In [1]: import numpy as np
         from scipy.optimize import minimize
         import scipy.stats as stats

In [2]: import pandas as pd
         tbl = pd.read_excel('mle.xlsx')
         tbl

Out[2]:
   Id  Y  X
0   1  2  1
1   2  6  4
2   3  7  5
3   4  9  6
4   5  15 9
```

Now we will see the application of maximum likelihood estimation for a regression equation. Before that, I have explained various theories. Now, we will take one example. I will explain how to find out or how to estimate the parameter of a regression equation using the principle called maximum likelihood estimation. The file name is, I have taken as MLE. I was importing the necessary libraries, import numpy as np.

You see that this is a new one from scipy.optimize import minimize. I am going to import a function that will minimize a function, import scipy.stats as stats. So I have imported the data. So this is the Y variable is a dependent variable. There is a 5 data set. There is X. X is the independent variable.

(Refer Slide Time: 11:58)

```

In [3]: import statsmodels.api as sm
x = tb1['X']
y = tb1['Y']
x2 = sm.add_constant(x)
mod1 = sm.OLS(y,x2)
mod1 = mod1.fit()
print(mod1.summary())

```

C:\Users\UVA\Anaconda\lib\site-packages\statsmodels\compat\pandas.py:56: FutureWarning: The `recreate` and `will be removed in a future version. Please use the pandas.tseries module instead.`
 from pandas.core import datetools

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.980								
Model:	OLS	Adj. R-squared:	0.973								
Method:	Least Squares	F-statistic:	145.9								
Date:	wed, 11 Sep 2019	Prob (F-statistic):	0.00012								
Time:	10:05:16	Log-Likelihood:	-4.5811								
No. Observations:	5	AIC:	13.16								
Df Residuals:	3	BIC:	12.38								
Df Model:	1										
Covariance Type:	nonrobust										
const	-0.2882	std err	0.755	t	-0.382	P> t	0.728	[0.025	0.975]		
X	1.6176		0.134		12.079		0.001		1.191		2.044
Omnibus:	nan	Durbin-Watson:	1.405								
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.551								
Skew:	0.089	Prob(JB):	0.759								
Kurtosis:	1.384	Cond. No.	12.5								

$$y = -0.2882 + 1.6176x_1$$

$$b_0 + b_1 x_1$$

$$\downarrow$$

$$b_0 + \beta_1 x_1$$

For this X and Y, I have constructed a regression equation. What is that regression equation? You see that by using our least square method, I have constructed a regression equation. So when I go for the least square method, the regression equation is $y = -0.2882 + 1.6176x_1$. I am explaining this portion to you. This estimation was, we know that this was the y intercept; this was b_1 .

So with the help of what we have done, we are going to predict $\beta_0 + \beta_1$ for x_1 coefficient. So this is the actual value. We know that this sample β_0 , b_0 can help to estimate the population β_0 . Similarly, the sample β_1 can be used to estimate the β_1 . That is for the population. So this was the answer when we were using the least square method. You see the method is least square method. Now we are going to use concept of maximum likelihood estimation, then we have to verify this answer; whether we are going to get the same answer.

(Refer Slide Time: 13:24)

$$b_0 = -0.2882 \text{ and } b_1 = 1.6176$$

This is the answer. We got the y intercept is -0.2882 and b1 x1 coefficient is 1.6176.

(Refer Slide Time: 13:32)

Parameter estimation by MLE

```
In [9]: e=modl2.resid  
  
In [10]: e  
Out[10]: 0    0.670588  
          1   -0.182353  
          2   -0.800000  
          3   -0.417647  
          4    0.729412  
         dtype: float64  
  
In [6]: hp.std(e)  
Out[6]: 0.6048820983804831
```

Another parameter which is required for maximum likelihood estimate is, you have to predict the standard deviation of the error variable. That is your error term. So for that, you can type e = modl2.residual, we get e; this is the error term. So we have to find the standard deviation of this error term. We got 0.06. This also we are going to predict. What we are going to predict? We are going to predict b0, b1 and this standard deviation of the error term using maximum likelihood estimation.

(Refer Slide Time: 14:09)

Parameter estimation by MLE

```
In [11]: import numpy as np
from scipy.optimize import minimize
import matplotlib.pyplot as plt

def lik(parameters):
    m = parameters[0]
    b = parameters[1]
    sigma = parameters[2]
    for i in np.arange(0, len(x)):
        y_exp = m * x + b
        l = (len(x)/2 * np.log(2 * np.pi) + len(x)/2 * np.log(sigma ** 2) + 1 /
             (2 * sigma ** 2) * sum((y - y_exp) ** 2))
    return l

x = np.array([1,4,5,6,9])
y = np.array([2,6,7,9,15])
lik_model = minimize(lik, np.array([2,2,2]), method='L-BFGS-B')

In [12]: lik_model
Out[12]: fun: 4.581084072762135
          hess_inv: <1x3 LbfgsInvHesProduct with dtype=float64>
          jac: array([1.24344979e-06, 2.84217094e-06, 1.33226763e-06])
         message: b'CONVERGENCE: NORM_OF_PROJECTED_GRADIENT_LT_PGTOL'
        nfev: 100
         nit: 17
       status: 0
      success: True
         x: array([-1.61764689, -0.28823426,  0.60488214])
```

This was the code for parameter estimation, for regression equations with the help of maximum likelihood estimation, import numpy as np from scipy.optimize import minimize import matplotlib.pyplot as plt. So I am defining a function that is going to give a likelihood function. So lik parameters, m is the slope b is the y-intercept, sigma is nothing but the standard deviation of the error term. So for i in np.arange 0 to all the value of x, we are going to find out y expected value is nothing but $mx + b$.

Then, this term is for estimating the log likelihood. So this term is nothing but when you go back previously, when you go back here, this is nothing but the whole equation. We are going to predict, that is why I used for loop. So for each I1, I2, I3, I4 up to I5, I will find out, then I will multiply it. That is why, this has come, this one. So this is nothing but that what I explained. Finally, the l will be returned. So this is our x value.

This x is taken from here, 1, 4, 5, 6, 9. This is nothing but this x value. 1, 4, 5, 6, 9; y value is 2, 6, 7, 9, 15. So like underscore model minimize, this is the function to minimize lik,np.array. So this is just, I am guessing the answer. What this first one says is slope, the second one says the y intercept, third one says the standard deviation of the error term. So I am going to use this method called; there are different methods.

I will show you what are the different methods for minimizing L underscore b of gs underscore b, this is one method. When you run it, see this is the answer 1.61 minus, what is this? This one is your slope. This is your x coefficient. This is your y-intercept. This is the standard deviation of your error term. When you go back, we will verify this. See the coefficient of x is 1.6176; here also getting 1.6176. The y-intercept is -0.288.

So here also getting y-intercept and other thing, we are getting the standard deviation of the error term is 0.604. So here also, see that this value also same. So what the point we are learning here is, the same problem can be done with the help of least square estimation method and maximum likelihood estimate method. In both the way, you will get the same answer. Now, we will take another example. This example, we have seen when I am teaching simple linear regression method.

(Refer Slide Time: 17:33)

Example 2

Example: Auto Sales

An Auto company periodically has a special week-long sale.

As part of the advertising campaign runs one or more television commercials during the weekend preceding the sale.

Data from a sample of 5 previous sales are shown on the next slide.

That you can recall that auto sales example. An auto company periodically has a special week long sale, as a part of advertising campaign runs one or more television commercials during the weekend preceding the sale. Data from the sample of 5 previous sales are shown in the next slide.

(Refer Slide Time: 17:54)

Example 2

Example: Auto Sales

<u>Number of TV Ads</u>	<u>Number of Cars Sold</u>
1	14
3	24
2	18
1	17
3	27

So what we have seen the number of TV ads is taken as independent variable; number of car sold is taken as dependent variable. So for this data set, first we will do a regression model with the help of least square estimation. Second, we will do with the help of maximum likelihood estimation. We will compare the answer; both will be the same. So first we will do, least square method.

(Refer Slide Time: 18:20)

Example 2

```
In [1]: import numpy as np
         from scipy.optimize import minimize
         import scipy.stats as stats

In [2]: import pandas as pd
        tbl = pd.read_excel('regcar.xlsx')
        tbl

Out[2]:
      TV Ads  car Sold
0       1       14
1       3       24
2       2       18
3       1       17
4       3       27
```

So I have imported the data. This was a TV ads and car sold.

(Refer Slide Time: 18:26)

```
In [3]: import statsmodels.api as sm
x=tbl[['TV Ads']]
y=tbl[['car Sold']]
x2 = sm.add_constant(x)
mod1 = sm.OLS(y,x2)
mod12 = mod1.fit()
print(mod12.summary())

```

OLS Regression Results

Dep. Variable:	car Sold	R-squared:	0.877		
Model:	OLS	Adj. R-squared:	0.836		
Method:	Least Squares	F-statistic:	21.43		
Date:	Wed, 11 Sep 2019	Prob (F-statistic):	0.0190		
Time:	11:11:23	Log-Likelihood:	-9.6687		
No. Observations:	5	AIC:	23.34		
Df Residuals:	3	BIC:	22.56		
Df Model:	1				
Covariance Type:	nonrobust				
coef	std err	t	P> t	[0.025	0.975]
const	10.0000	2.366	4.226	0.024	2.469 17.531
TV Ads	5.0000	1.080	4.629	0.019	1.563 8.437

.....

Omnibus:	nan	Durbin-Watson:	1.214
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.674
Skew:	0.256	Prob(JB):	0.714
Kurtosis:	1.276	Cond. no.	6.33

When I do, you see there is a OLS, ordinary least square method, we are getting this is the answer. What is that this answer? $Y = 10 + 5 \text{ TV ads}$. Here, you can say x_1 is TV ads. Now for this term, we will find out what is the error term.

(Refer Slide Time: 18:47)

$b_0 = 10$ and $b_1 = 5$

```
In [4]: e=mod12.resid
```



```
In [5]: e
```

```
Out[5]: 0    -1.0
       1    -1.0
       2    -2.0
       3     2.0
       4     2.0
      dtype: float64
```



```
In [9]: np.std(e)
```

```
Out[9]: 1.6733200530681507
```

See for finding the error term, see the b_0 is 10, b_1 is 5. To find the error term, I am going to save in the object called $e = \text{mod12.residual}$ e . So this was the error term. So if I find the standard deviation of this error term, we are getting 1.67. So we are going to predict this standard deviation of the error term and your y intercept and the coefficient of x with the help of maximum likelihood estimation. So there also, we will get the same answer.

(Refer Slide Time: 19:20)

```
In [14]: M import numpy as np
         from scipy.optimize import minimize
         import matplotlib.pyplot as plt

def lik(parameters):
    m = parameters[0]
    b = parameters[1]
    sigma = parameters[2]
    for i in np.arange(0, len(x)):
        y_exp = m * x + b
    L = (len(x)/2 * np.log(2 * np.pi) + len(x)/2 * np.log(sigma ** 2) + 1 /
         (2 * sigma ** 2) * sum((y - y_exp) ** 2))
    return L

x = np.array([1,3,2,1,3]) ✓
y = np.array([14,24,18,17,27]) ✓ ✓ ✓
lik_model = minimize(lik, np.array([2,2,2]), method='Nelder-Mead')

In [15]: M lik_model
Out[15]: final_simplex: (array([[ 5.00000631, 10.00000822, 1.67332132],
```

What we have done? The same thing, because already I have defined the function, it will be easy for me, just you replace the various parameters. If 0th index is m, 1 is b, 2 is sigma, so this is our likelihood function. So this is my x value. This is my y value. This is the function to minimize. I will run this program after a few minutes. This is just I have taken the screenshot of the python program for your explanation purpose.

You see that the final value, you look at this, this is your slope is 5, because this one. Second one is the y intercept is 10. See the standard deviation of the error term is 1.67. It is exactly what we have done using the least square method. Now I will go to Python environment. I will run and will explain and one more thing you have to remember this is my guessed value 2, 2, 2. While running this program, I am going to change this values, still we are going to get the same answer. This is just, I am guessing the value. You can give any value. At the end, you will get the same answer. Now we will go to the Python environment. We will do the program.

(Refer Slide Time: 20:45)

```

In [ ]: M import numpy as np
         from scipy.optimize import minimize
         import scipy.stats as stats
         import matplotlib.pyplot as plt

In [ ]: M import pandas as pd
         tbl = pd.read_excel('MLE.xlsx')
         tbl

In [ ]: M import statsmodels.api as sm
         x=tbl['X']
         y=tbl['Y']
         x2 = sm.add_constant(x)
         mod1 = sm.OLS(y,x2)
         mod12 = mod1.fit()
         print(mod12.summary())

b0 = -0.2882 and b1 = 1.6176

In [ ]: M e=mod12.resid
         e

```

I have explained how to use maximum likelihood estimation to predict the parameter of a regression equation. I have shown the screenshot of the program in my presentation. Now using Python environment, we will run this code. I will explain how it is working. I have imported the necessary libraries, then I have stored my data in the file called MLE. So I have displayed this data. This data says, y is the dependent variable, x is independent variable. For this data set, we are going to construct a regression equation using least square method.

(Refer Slide Time: 21:27)

```

In [ ]: M mod12 = mod12.fit()
         print(mod12.summary())

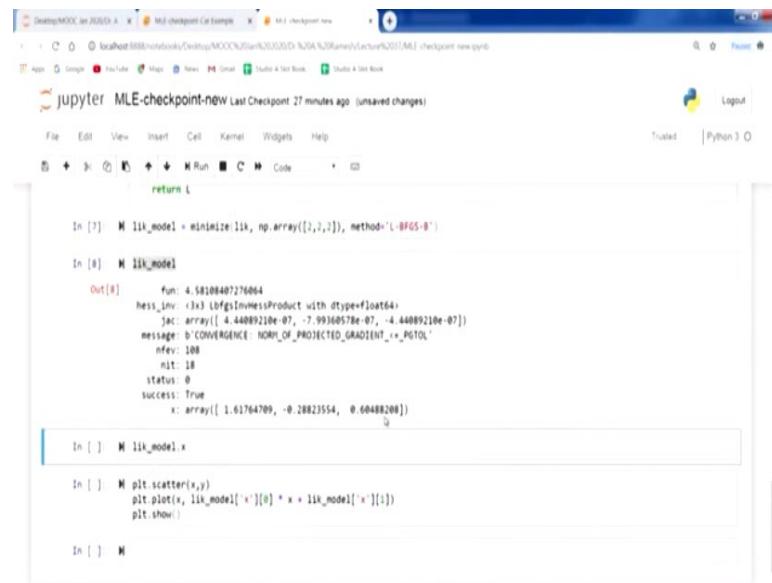
OLS Regression Results
=====
Dep. Variable: Y R-squared: 0.980
Model: OLS Adj. R-squared: 0.973
Method: Least Squares F-statistic: 145.9
Date: Thu, 12 Sep 2019 Prob (F-statistic): 0.00122
Time: 16:44:19 Log-Likelihood: -4.5811
No. Observations: 5 AIC: 13.16
Df Residuals: 3 BIC: 12.38
Df Model: 1
Covariance Type: nonrobust
=====
            coef std err      t      P>|t|    [0.025   0.975]
-----
const    -0.2882    0.755    -0.382    0.728    -2.692     2.115
X         1.6176    0.134    12.079    0.001    1.191     2.044
-----
Omnibus: nan Durbin-Watson: 1.405
Prob(Omnibus): nan Jarque-Bera (JB): 0.551
Skew: 0.089 Prob(JB): 0.759
Kurtosis: 1.389 Cond. No. 12.5
=====

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```

This was the output of least square method regression model. Here the y intercept is – 0.2882. The coefficient of x is 1.6176. So how can we write the regression equation? $Y = -0.2882 + 1.6176x$.

(Refer Slide Time: 21:49)



The screenshot shows a Jupyter Notebook window titled "jupyter MLE-checkpoint-new". The notebook has two cells:

- In [7]:** `M lik_model = minimize(lik, np.array([1,1,1]), method='L-BFGS-B')`
- In [8]:** `Out[8]:`
Detailed output from the optimization process, including:
 - fun: 4.58108407276064
 - hess_inv: (3,3) LbfgsInversesProduct with dtype='float64'
 - jac: array([-4.44089210e-07, -7.99360578e-07, -4.44089210e-07])
 - message: b'CONVERGENCE: NORM_OF_PROJECTED_GRADIENT_<= PGtol'
 - nfev: 108
 - nit: 18
 - status: 0
 - success: True
 - x: array([1.61764709, -0.28823554, 0.60488208])

Next one, this is y intercept. This is the coefficient of x_1 . Next, we are going to find out the error by using this command. That is `dot resid`. For this error term, we have to find the standard deviation of the error term. The standard deviation of the error term is 0.60488. Now there are three things which we have done. We have found the coefficient y intercept and the error term. Now, by using the concept of maximum likelihood estimation, you will verify this answer.

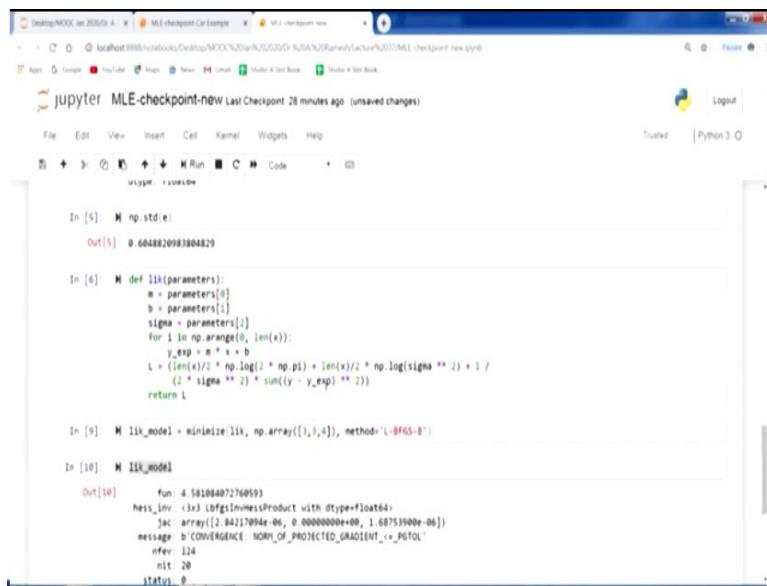
Whether we are getting same standard deviation of the error, same coefficient, and same y intercept. I have defined a function; the function name is `lik`. So I have called the slope and y intercept and σ in `np.arange(0, ln x)`. I have predicted what is the y expected value by substituting different x value $mx + b$. So this l is nothing but the likelihood function. This likelihood function, I have explained this formula in my presentation.

So I am going to run this for all value of x_1 , then I am going to return the value l . This function is going to return the value l . So I am going to minimize the likelihood function, because the error has to be minimized. So this 2, 2, 2 these values randomly I am guessing, what will be the parameter, that is m, b , and the standard deviation of the error term. So I am going to display this model. So this model says that my slope is 1.617.

You see that when you do the OLS method also, the slope is 1.6176. The constant, see here constant is – 0.288. In OLS method also the constant is – 0.2882. Next we predicted the standard deviation of the error term, that is 0.604. Look at here also, we got the standard deviation of the error term is 0.604. Now what I am going to do? I am going to change this value. For example, 2 I am going to give 3. This I will give 4. Let us see what value, we are going to get.

Again, we are getting, there is no change in the answer. So this value, this np.array, this is our guessing value for our parameter. So at the end, we are getting the same answer. This is our example number 1. Now I will go to another example.

(Refer Slide Time: 24:33)



```
In [5]: M np.std(e)
Out[5]: 0.6048820983804829

In [6]: M def llik(parameters):
    a = parameters[0]
    b = parameters[1]
    sigma = parameters[2]
    for i in np.arange(0, len(x)):
        y_exp = a * x + b
        L = (len(x)/2 * np.log(2 * np.pi) + len(x)/2 * np.log(sigma**2)) + 1 /
            (2 * sigma**2) * sum((y - y_exp)**2)
    return L

In [9]: M llik_model = minimize(llik, np.array([1,1]), method='L-BFGS-B')

In [10]: M llik_model
Out[10]:
      fun: 4.581884072760599
      hess_inv: <3x3 LbfgsInvHesProduct with dtype=<float64>
      jac: array([2.84217994e-06, 0.00000000e+00, 1.68753900e-06])
     message: b'CONVERGENCE: NORM_OF_PROJECTED_GRADIENT_<=PGTOL'
     nfev: 124
     nit: 20
    status: 0
```

This example, when I am explaining linear regression, I solved this problem with the help of simple linear regression by using least square method. I will clear my output. Here also, we are going to do the same thing, what we have done for previous problem. We are going to predict the regression model using OLS, then we are going to check that answer with the help of our maximum likelihood estimation methods.

So I am importing the library, then I am calling the data. This is the data set. The TV ad is the independent variable, car sold is dependent variable. Now I am going to construct a regression equation by using OLS, ordinary least square method. This answer is, the 10 is the constant, 5 is

the coefficient of TV ads. So we can write $y = 10 + 5$ TV ads. Next, I am going to find out the error term.

(Refer Slide Time: 25:36)

```

In [6]: np.std(e)
Out[6]: 1.6733200530681511

In [7]: def lik(parameters):
    m = parameters[0]
    b = parameters[1]
    sigma = parameters[2]
    for i in np.arange(0, len(x)):
        y_exp = m * x + b
        L = (len(x)/2 * np.log(2 * np.pi) + len(x)/2 * np.log(sigma ** 2)) + i /
            (2 * sigma ** 2) * sum((y - y_exp) ** 2))
    return L

In [8]: lik_model = minimize(lik, np.array([5,5,5]), method='Nelder-Mead')

In [9]: lik_model.x
Out[9]: array([ 4.99997913, 10.00000001, 1.67332791])

In [ ]: plt.scatter(x,y)
plt.plot(x, lik_model['x'][0] * x + lik_model['x'][1])
plt.show()

In [ ]: minimize?

```

So the error term is, this is the error term. Now I am going to find out the standard deviation of the error term. The standard deviation of the error term is 1.67. So now these parameter, which I have got with the help of least square method, I am going to get the same answer with the help of maximum likelihood estimate. I am calling the same function. So what is more important here, with this function is $L = \text{length } x \text{ divided by } 2, \text{ this star } np.\log 2 \text{ star } np.pi$.

This I have explained in my slide, when there is a normal distribution, if you want to find out the parameter of that, we have to use this formula. That I have explained in my class. You can refer my previous slides there. This 2, 2, 2 is the guessing values. For example, I am going to do 5. I am going to change this number to 5. Now let us run it. You see that the 4.99 actually our answer is 5; we got 4.99, approximately correct.

The y-intercept is 10, here also we got y intercept is 10 and the standard deviation of the error term is 1.67. So here also, we are getting 1.67. So this way, we have verified with the help of this Python program that whatever answer which you get with the help of OLS is the same as maximum likelihood estimation. Because this maximum likelihood estimation method for predicting the population parameter is so generic and most of the software packages, they follow

this maximum likelihood estimate for predicting the population parameter. As I told you, there are different ways to minimize. Suppose, if you want to check the different methods, simply minimize, put this question mark, you will get different methods.

(Refer Slide Time: 27:44)

```

y_exp = x * b
L = (len(x)/2) * np.log(2 * np.pi) + len(x)/2 * np.log(sigma ** 2) + 1 / (2 * sigma ** 2) * sum((y - y_exp) ** 2)
return L

In [8]: M lik_model = minimize(lik, np.array([5,5]), method='Nelder-Mead')

In [9]: M lik_model.x
Out[9]: array([ 4.99997913, 10.00000001, 1.67332791])

```

derivatives ('fun', 'jac' and 'hess' functions).
method : str or callable, optional
Type of solver. Should be one of

- 'Nelder-Mead' : ref '(see here) <optimize.minimize.neldermead'
- 'Powell' : ref '(see here) <optimize.minimize_powell'
- 'CG' : ref '(see here) <optimize.minimize_cg'
- 'BFGS' : ref '(see here) <optimize.minimize_bfgs'
- 'Newton-CG' : ref '(see here) <optimize.minimize_newtoncg'
- 'L-BFGS-B' : ref '(see here) <optimize.minimize_lbfgsb'
- 'TNC' : ref '(see here) <optimize.minimize_tnc'
- 'COV_NL' : ref '(see here) <optimize.minimize_cobyla'
- 'SLSQP' : ref '(see here) <optimize.minimize_slsqp'
- 'trust-constr' : ref '(see here) <optimize_trustconstr'
- 'dogleg' : ref '(see here) <optimize.minimize_dogleg'
- 'trust-nocg' : ref '(see here) <optimize.minimize_trustnogc'
- 'trust-ksa' : ref '(see here) <optimize.minimize_trustksa'

One method is see that Nelder-Mead, Powell, CG, BFGS, Newton CG, there are different methods. So what we can do, here there is a method. You can change some other value, then we will get the same answer. In this class, I have given an example how to use maximum likelihood estimation method for predicting the population parameter of a regression equation. I have explained the theory. I have taken two examples for that.

For the two examples also, I have concluded that you can use OLS method, that is ordinary least square method and maximum likelihood method. In both the way, you will get the same answer. Then I have explained how to do the coding and how to run and get the answer using Python, because this class is based for the next class, which I am going to take, logistic regression, because the logistic regression, the method which you are going to use to predict the population parameter is maximum likelihood estimation. In the next class, by applying this principle of MLE, maximum likelihood estimation, we will use and estimate the population parameter of logistic regression.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture 38
Logistic Regression - I

In this class, we are going to new topic that is the logistic regression. I am going to explain, when you go for logistic regression over linear regression and see generally when we are doing linear regression, both independent variable and dependent variable are continuous. When in independent variable there is a categorical data, we have used dummy variable regression. There may be a chance even in the dependent variable, there may be some categorical variable.

In that case, we should go for logistic regression. I will explain this logistic regression with the help of example, then I will interpret our Python output, at the end I will explain the theory behind this logistic regression.

(Refer Slide Time: 01:14)

Agenda

- Building Logistic regression Model
- Python Demo on Logistic Regression

The class agenda is that we will build a logistic regression model, then I will do the demo for logistic regression model. We will see the application of logistic regression.

(Refer Slide Time: 01:24)

Application

$$y = a + b_1x_1 + b_2x_2$$

- In many regression applications the dependent variable may only assume two discrete values.
- For instance, a bank might like to develop an estimated regression equation for predicting whether a person will be approved for a credit card or not
- The dependent variable can be coded as $y = 1$ if the bank approves the request for a credit card and $y = 0$ if the bank rejects the request for a credit card.
- Using logistic regression we can estimate the probability that the bank will approve the request for a credit card given a particular set of values for the chosen independent variables.

In many regression applications, the dependent variable may only assume 2 discrete variables. For example, in the linear regression also the dependent variable was a continuous variable, independent variable also continuous, but some time what may happen, the dependent variable may be discrete values. For example, gender. For example, good or bad or success or failure. So the y value, so what is happening? See in a linear regression, $y = a + b_1x_1 + b_2x_2$.

If x_1, x_2 are independent variable, y is a dependent variable, in the x_1 if there is any categorical data, we should go for a dummy variable regression. Sometime in the dependent variable, there may be some categorical data. For example, 0, 1, it may be gender. It may be quality of product, good or bad, whether a person will buy the product or not buy the product. Whenever there is two options, it may be categorical. That time, we should go for a logistic regression.

For instance, a bank might like to develop your estimated regression equation for predicting whether a person will be approved for a credit card or not. Here the y , the dependent variable is a person will be approved for getting credit card or not. We have two possibility, then you should go for logistic regression. The dependent variable can be coded as $y = 1$, if the bank approves the request for a credit card and $y = 0$ if the bank rejects request for a credit card.

Using logistic regression, we can estimate the probability that the bank will approve the request of a credit card given a particular set of values for the chosen independent variable. This may be

applicable, when you go for applying for loan. Whether this person will repay the loan or not; because there are two possibilities, so that also we can use logistic regression. Somebody applying for some jobs whether he will get the job or he will not get the job.

For that purpose, you can go for logistic regression. In your context, we can say whether you will get the placement or not. Here it is only two possibilities, a person may get or may not get the placement. So that case, we can have different independent variables. So what kind of independent variables will help you to get the placement, that kind of problem can be solved with the help of this logistic regression model.

(Refer Slide Time: 04:06)

Example

- Let us consider an application of logistic regression involving a direct mail promotion being used by [Simmons Stores](#).
- Simmons owns and operates a national chain of women's apparel stores.
- Five thousand copies of an expensive four-color sales catalog have been printed, and each catalog includes a coupon that provides a \$50 discount on purchases of \$200 or more.
- The catalogs are expensive and Simmons [would like to send them to only those customers who have the highest probability of using the coupon](#).

Sources: Statistics for Business and Economics, 11th Edition by David R. Anderson (Author), Dennis J. Sweeney (Author), Thomas A. Williams (Author)

Let us take one example. This example is taken from the book Statistics for Business and Economics, 11th Edition by David Anderson, Dennis Sweeney and Thomas Williams. I will suggest you, this book is excellent book to understand the concepts. This problem also, which I have taken in this lecture, is from this book. Let us consider an application of logistic regression involving direct mail promotion being used by Simmons Stores. The store name is Simmon.

So they are going for a promotion. Simmon owns and operates a national chain of woman's apparel stores. 5000 copies of expensive four colour sales catalog have been printed and each catalog includes a coupon that provides a \$50 discount on purchase of \$200 or more. The catalogs are expensive and Simmons would like to send them to only those customers, who have

the highest probability of using the coupon. Now we have to identify, for what kind of customers we have to target, so that they will use the coupon.

(Refer Slide Time: 05:22)

Variables

- Management thinks that **annual spending** at Simmons Stores and whether a **customer has a Simmons credit card** are two variables that might be helpful in predicting whether a customer who receives the catalog will use the coupon.
- Simmons conducted a pilot study using a random sample of 50 Simmons credit card customers and 50 other customers who do not have a Simmons credit card.
- Simmons sent the catalog to each of the 100 customers selected.
- At the end of a test period, Simmons noted whether the customer used the coupon or not?

What are the variables in these problems? The management thinks that the annual spending is one of the variable at Simmon stores and whether a customer has see a Simmon credit card are two variables that might be helpful in predicting whether a customer who receives the catalog will use the coupon. So there are two independent variables. One is annual spending. Second one whether the person having some Simmon's credit card or not.

Simmons conducted a pilot study using a random sample of 50 Simmons' credit card customers and 50 other customers who do not have the Simmon credit card. Simmons sent to the catalog to each of the hundred customers selected. At the end of the test period, Simmons noted whether the customers used the coupon or not. By using this data set, they are going to construct a regression equation model, so that they can target to whom this catalog can be sent, so that they will use this coupon, so that the sales will increase.

(Refer Slide Time: 06:29)

Data (10 customer out of 100)

Customer	x_1 Spending	x_2 Card	y Coupon
1	2.291	1	0
2	3.215	1	0
3	2.135	1	0
4	3.924	0	0
5	2.528	1	0
6	2.473	0	1
7	2.384	0	0
8	7.076	0	0
9	1.182	1	1
10	3.345	0	0

This is a data set. I have shown for 10 customers, but there are 100 dataset; 50 dataset people are those who are not having the credit card, the remaining 50 are those who are having the credit card. So spending is one of the independent variable. Having or not having, for example 1 means having the credit card, 0 means not having the credit card. Here the coupon also 0 means they have not used the coupon, 1 means they have used the coupon.

So the coupon this variable, this is going to be our dependent variable. This is one independent variable x_1 , this is another independent variable x_2 . You see that the x_2 variable is a categorical variable right. Actually in case, there are different levels. We have to convert into a dummy variable, then you have to run the analysis, but in this problem directly it is given, whether the person is having the credit card or not having the credit card.

(Refer Slide Time: 07:30)

Explanation of Variables

- The amount each customer spent last year at Simmons is shown in thousands of dollars and the credit card information has been coded as 1 if the customer has a Simmons credit card and 0 if not.
- In the Coupon column, a 1 is recorded if the sampled customer used the coupon and 0 if not.

Now we will go for what is the explanation of variables. The amount of each customer spent last year at Simmons is shown in 1000s of dollars and the credit card information has been coded as 1 if the customer has the Simmon credit card, 0 if not. So two variables, one is how much spent the last year, whether the person having the credit card or not. If the person is having credit card 1, otherwise it is 0. In the coupon column which is dependent variable 1 is recorded if the sampled customer used the coupon, 0 means if not.

(Refer Slide Time: 08:10)

Logistic Regression Equation

- If the two values of the dependent variable y are coded as 0 or 1, the value of $E(y)$ in equation given below provides the probability that $y = 1$ given a particular set of values for the independent variables x_1, x_2, \dots, x_p .

LOGISTIC REGRESSION EQUATION

$$E(y) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}$$

First we will go for what is the logistic regression equation. If the two values of the dependent variable y are coded as 0 or 1, the value of expected y in equation given below provides the probability that $y = 1$ given you a particular set of values for the independent variable x_1, x_2, x_p .

So logistic regression equation is expected value of $y = e$ to the power ($\beta_0 + \beta_1 x_1 + \beta_2 x_2$ up to β_p). There are p independent variables, divided by $(1 + e$ to the power ($\beta_0 + \beta_1 x_1 + \beta_2 x_2$ and so on up to $\beta_p x_p$)). So this y is we are going to predict y . It is going to be 0 or 1.

(Refer Slide Time: 09:04)

Logistic Regression Equation

- Because of the interpretation of $E(y)$ as a probability, the **logistic regression equation** is often written as follows

$$E(y) = P(y = 1 | x_1, x_2, \dots, x_p)$$

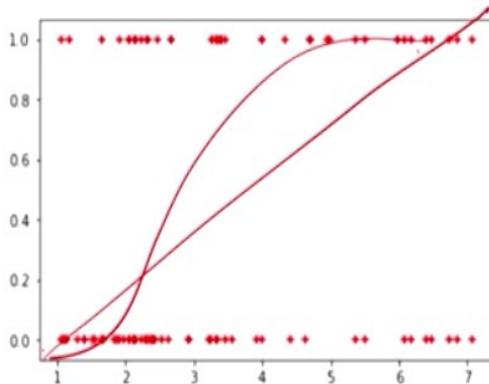
You may ask the question, why not we use simple linear regression equation here? Because the simple linear regression equation cannot be used for this problem, because there are two possibilities. We have assumed that when you plot this data set, I will tell you the next slide that is there. One assumption is that the error term in a simple linear regression should follow a normal distribution, but here the Y variable is only two possibilities. So that will follow binomial distribution and the error term of a logistic regression will follow binomial distribution.

So you cannot use your simple linear regression, whenever there is a y -value is categorical variable. Because of the interpretation of expected y is a probability of logistic regression equation is often written as expected value of $y = P(y = 1 \text{ given } x_1, x_2 \text{ up to } x_p)$. So we are going to find out the expected value of y .

(Refer Slide Time: 10:08)

```
In [15]: plt.scatter(df.Spending,df.Coupon,marker='+',color= 'red')
```

```
Out[15]: <matplotlib.collections.PathCollection at 0x2a1b5b73c50>
```

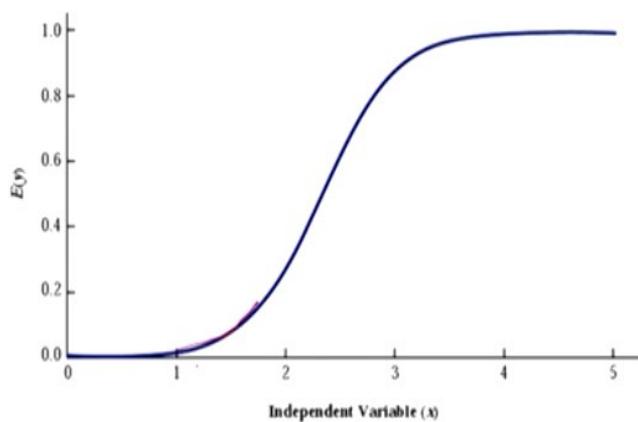


This was the example that why we cannot use linear regression. So what will happen, here in no way you cannot construct any linearity, because here the x variable is spending. Spending is a continuous variable. The y variable is a person has used coupon or not. So what is happening? Whenever their income is low, they also used the coupon. When the incomes are more, that time also they have used the coupon.

So you cannot construct, for this data set, when you fit to this kind of linear regression, there is no meaning for that. So one way to fit the line for this kind of data is the S-shaped curve that I will show you in the next slide. It will become this way, that I will show in the next slide.

(Refer Slide Time: 11:01)

Logistic regression equation for β_0 and β_1



Yeah, this one assume that there are independent variable is there. The range of independent variable up to say 0 to 5, the expected value is nothing but the probability. You see that when this $x = 3$, it is getting the maximum value between 3 and 4. Whenever the value of x is between 3 and 4, there is a higher chances that expected value of y will be 1. When it is below 2, when the x value is below 2, there is a higher chance that the expected value will become 0.

So there are two possibilities and the rate of change also you see here, the rate of change also low between 1 and 2, but between 2 & 3 the rate of change is high, but between 3 and 4 the rate of change is low. So this is an S-shaped curve for a logistic equation. So what we are understanding here, when $x = 3$ whenever the value of x is more than 3, there is a more chance the value of expected value of y will be 1. When it goes below 1 or below 2, there is more chance that the expected value of y will be 0. When you go right hand side, there is a more chance that the expected value of y becomes 1.

(Refer Slide Time: 12:24)

Logistic regression equation for β_0 and β_1

- Note that the graph is S-shaped.
- The value of $E(y)$ ranges from 0 to 1, with the value of $E(y)$ gradually approaching 1 as the value of x becomes larger and the value of $E(y)$ approaching 0 as the value of x becomes smaller.
- Note also that the values of $E(y)$, representing probability, increase fairly rapidly as x increases from 2 to 3.
- The fact that the values of $E(y)$ range from 0 to 1 and that the curve is S-shaped makes equation (slide no.11) ideally suited to model the probability the dependent variable is equal to 1.

Now I will explain what is that previous curve? Note that the graph is S-shaped. The value of expected y range from 0 to 1, that is in x axis, with the value of expected value of y gradually approaching 1 as the value of x becomes larger and the value of expected value of y approaching 0 as the value of x become smaller. Note also that the value of expected y representing probability increase fairly rapidly as x increases from 2 to 3 after that it becoming constant.

The fact that the value of expected value of y range from 0 to 1 and that the curve S-shaped makes the equation, the previous slide this shape ideally suited to model the probability that dependent variable is equal to 1.

(Refer Slide Time: 13:31)

Estimating the Logistic Regression Equation

- In simple linear and multiple regression the least squares method is used to compute b_0, b_1, \dots, b_p as estimates of the model parameters (b_0, b_1, \dots, b_p).
- The nonlinear form of the logistic regression equation makes the method of computing estimates more complex MLE
- We will use computer software to provide the estimates.
- The estimated logistic regression equation is

ESTIMATED LOGISTIC REGRESSION EQUATION

$$\hat{y} = \text{estimate of } P(y = 1 | x_1, x_2, \dots, x_p) = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p}}$$

Here, \hat{y} provides an estimate of the probability that $y = 1$, given a particular set of values for the independent variables.

Now we will go for estimation of logistic regression equation. In a simple linear and multiple regression, the least square method is used to compute b_0, b_1 up to b_p as estimates of the model parameter. What is the model parameter? Zero, this is beta 0, beta 0, beta 1 up to beta p. So what you have done? With the help of this sample parameter, we have estimated the population parameter, that is beta 0, beta 1 and beta p.

But the previous equation that is the logistic equation is non linear. So the non linear form of the logistic regression equation makes the method of computing estimates more complex. So what we are going to do? That is why in the previous class as I explained, whenever there is a non linear form of equation, instead of using that OLS method, you have to use your maximum likelihood estimation method, MLE to predict the population parameter.

So all software packages follow the concept of maximum likelihood estimation, I have explained the previous class to get, to predict the population parameter with the help of sample parameter. We will use computer software, the Python to provide the estimate. At the end of the class, I will

show you that. The estimated logistic regression equation is \hat{y} is nothing but p of $y = 1$ for different x_1, x_2, x_p equal to e to the power (b_0, b_1, \dots). This b_0, b_1 is these sample parameter.

Divided by $(1 + e^{(b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p)})$. Here \hat{y} provides an estimate of the probability that $y = 1$ given a particular set of values of the independent variables. The \hat{y} is the probability, that probability will tell us how much chance the p of $y = 1$. If it is a higher probability that p of $y = 1$. If \hat{y} is low, you will get a lower probability.

(Refer Slide Time: 15:57)

Python Code for Logistic Regression

```
In [12]: x = df[['Card', 'Spending']]
y = df['Coupon']

import statsmodels.api as sm
x1= sm.add_constant(x)
logit_model=sm.logit(y,x1)
result=logit_model.fit()
print(result.summary2())

Optimization terminated successfully.
  Current function value: 0.604869
    Iterations 5
      Results: logit
-----
Model:          Logit      No. Iterations:  5.0000
Dependent Variable: Coupon   Pseudo R-squared: 0.101
Date: 2019-09-11 12:54 AIC: 126.9739
No. Observations: 100 BIC: 134.7894
Df Model: 2 Log-likelihood: -60.487
Df Residuals: 97 LL-Null: -67.301
Converged: 1.0000 Scale: 1.0000
-----
          Coef.  Std.Err.      z     P>|z|    [0.025  0.975]
-----
const  -2.1464  0.5772 -3.7183  0.0002 -3.2778 -1.9150
Card   1.0987  0.4447  2.4707  0.0135  0.2271  1.9703
Spending 0.3416  0.1287  2.6551  0.0079  0.0894  0.5938
-----
```

6

I have brought the screenshot of the logistic regression. There are two independent variables. One is card and spending. There y is a dependent variable. So I am going to use a constant $x_1 = sm.add_constant$. Here you see that we are going to use Logit model. Logit underscore model equal to $sm.logit(y, x1)$. result equal to $logit_model.fit()$. Print the $result.summary2()$, then you will get this output. So look at this.

This is the constant is -2.14. This coefficient of card is 1.0987. The coefficient of spending is 0.3416. See there are 100 observations. I have shown only in my previous slides, only 10 observation only for understanding purpose. The model is Logit model and there are pseudo R square is 0.101. There is AIC. There is a log likelihood and log likelihood when the variable is not there. That is a log likelihood underscore null.

This we will use to find out the G statistics. I will tell you later. Then this is standard error of this regression coefficient. The z value, it is called wald statistic, we can say WALD statistics. That is nothing but the coefficient 1.0987 divided by 0.4447, you will get this one. This was the p-value. There are two things you have to understand before interpreting the answer. One is we have to look at the G statistics. In the coming slides I will explain what is the G statistic.

That G statistics is equivalent to F statistics of our linear regression. What we have done? The F statistics in the linear regression is helping to test the overall model and the t statistics in the linear regression is used to check the significance of an individual independent variable. The same way here the G statistics is to test the significance of overall logistic regression model. Here the z that is the WALD statistics is used to test the significance of individual the corresponding p value, is used to test the significance of individual independent variable. That is meant each independent variable. I will go further, then I will explain what is the meaning of that.

(Refer Slide Time: 18:50)

Variables

$$y = \begin{cases} 0 & \text{if the customer did not use the coupon} \\ 1 & \text{if the customer used the coupon} \end{cases}$$

x_1 = annual spending at Simmons Stores (\$1000s)

$$x_2 = \begin{cases} 0 & \text{if the customer does not have a Simmons credit card} \\ 1 & \text{if the customer has a Simmons credit card} \end{cases}$$

$$E(y) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}$$

$$\hat{y} = \frac{e^{b_0 + b_1 x_1 + b_2 x_2}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2}} = \frac{e^{-2.14637 + 0.341643x_1 + 1.09873x_2}}{1 + e^{-2.14637 + 0.341643x_1 + 1.09873x_2}}$$

So what are the variables? We have taken y, y can have two possibilities 0 if the customer did not use the coupon, 1 if the customer used the coupon. x_1 is the annual spending at Simmon stores that is in terms of 1000, then x_2 is a categorical variable. Categorical variable is 0 if the customer does not have the Simmon credit card, 1 if the customer has the Simmon credit card. So we know that the expected value of $y = e$ to the power ($\beta_0 + \beta_1 x_1$).

This was for the population, but this can be done with the help of \hat{y} that is $e^{b_0 + \beta_1 x_1 + \beta_2 x_2}$ divided by $(1 + e^{b_0 + \beta_1 x_1 + \beta_2 x_2})$. So this was the sample statistic. So from the previous output, what is the b_0 here? See b_0 is -2.1464. We got -2.1464. Now in our problem, the x_1 spending, how much the customer has spent in the last time last year, see that is taken our x_1 variable. Here x_2 , is the person is processing the card or not.

So the constant is -2.1464. So here x_1 is, the coefficient of x_1 is 0.34164. We got this one, 0.3416 and the coefficient of x_2 is 1.0987. So we are getting 1.0987. So this was in the numerator, then $1 + e$ to the power the same value in the denominator.

(Refer Slide Time: 20:43)

Managerial Use

- $P(y = 1/x_1 = 2, x_2 = 0) = .1880$

$$\hat{y} = \frac{e^{-2.14637 + 0.341643(2) + 1.09873(0)}}{1 + e^{-2.14637 + 0.341643(2) + 1.09873(0)}} = \frac{e^{-1.4631}}{1 + e^{-1.4631}} = \frac{.2315}{1.2315} = 0.1880$$

- $P(y = 1/x_1 = 2, / x_2 = 1) = .4099$

$$\hat{y} = \frac{e^{-2.14637 + 0.341643(2) + 1.09873(1)}}{1 + e^{-2.14637 + 0.341643(2) + 1.09873(1)}} = \frac{e^{-0.3644}}{1 + e^{-0.3644}} = \frac{.6946}{1.6946} = 0.4099$$

- Probabilities indicate that for customers with annual spending of \$2000 the presence of a Simmons credit card increases the probability of using the coupon

We have got the output of logistic regression equation model. We will look at interpret and we will see the managerial use of that. How to interpret this one? For example, when $y = 1$, $x_1 = 2$, $x_2 = 0$. What is the meaning? Suppose the person's income is \$2,000, is not having the credit card. When you substitute in our estimated regression equation, substitute $x_1 = 2$, $x_2 = 0$, both the numerator and denominator. When you simplify, we are getting 0.1880.

What is the meaning is that a person is having or not having credit card and having the expenditure of \$2,000 that probability of that fellow to use that coupon is 0.1880. The same case instead of $x_2 = 0$, I am going to see the interpretation $x_2 = 1$, that means what? A person having

the credit card, spending \$2,000, what is the probability that that person will use the coupon? So you substitute $x_1 = 2$. Previously, we substituted $x_2 = 0$, now substituted $x_2 = 1$.

So when you simplify this, we are getting 0.4099. So what has happened? A person having the credit card is having the more possibility of that is the probability of $y = 1$, becomes higher. That means, there is a more chance that a person having the credit card will use the coupon. So probabilities indicates that the customers with the annual spending of \$2,000, the presence of a Simmon credit card increases the probability of using the coupon.

How it is increasing? You see that this much. This is the probability of same income, but not having the credit card. This is the probability of having credit card. So what is happening? The probability is increased if the person is processing the credit card.

(Refer Slide Time: 22:45)

Managerial Use

- It appears that the probability of using the coupon is much higher for customers with a Simmons credit card.

		Annual Spending						
		\$1000	\$2000	\$3000	\$4000	\$5000	\$6000	\$7000
Credit Card	Yes	0.3305	0.4099	0.4943	0.5791	0.6594	0.7315	0.7931
	No	0.1413	0.1880	0.2457	0.3144	0.3922	0.4759	0.5610

Like that, for different possibilities, so previously we have explained only this portion. Now $x_1 = 1$, that means \$1,000 $x_2 = 1$, you will get this probability, then $x_1 = 1$, $x_2 = 0$, you will get this probability. Like that we have extended for up to \$7,000. When you look at this figure, you see that when you compare the probability, the person having the credit card there is a more chances are that that fellow will use the coupon. If the person is not having credit card, there is a lesser chance to use the credit card. That is one interpretation.

(Refer Slide Time: 23:34)

Testing for Significance

t, f

$$H_0: \beta_1 = \beta_2 = 0$$

H_a : One or both of the parameters is not equal to zero

Before interpreting, we have to test whether the coefficient are significant or not, because the equation which we have can constructed is only for the population. The same thing also we have done our linear regression equation. The linear regression equation we have used a t-test and F test to predict the significance of the independent variable. So what is the null hypothesis? Beta 1 = beta 2 = 0, so one or both of the parameter is not equal to 0.

(Refer Slide Time: 24:07)

G Statistics

- The test for overall significance is based upon the value of a G test statistic.
- If the null hypothesis is true, the sampling distribution of G follows a chi-square distribution with degrees of freedom equal to the number of independent variables in the model.

F

Now we will go for G statistics. The test for overall significance is based upon the value of G test statistics. This is equivalent to F test statistics in our linear regression model. If the null hypothesis is true the sampling distribution of G follows a chi-square distribution with the degrees of freedom equal to the number of independent variable in the model. In our problem,

the number of independent variable is 2. So the degrees of freedom is 2. If there is only one independent variable, the degrees of freedom for G statistics is 1.

(Refer Slide Time: 24:43)

```
In [12]: x = df[['Card','Spending']]
y = df['Coupon']

import statsmodels.api as sm
x1= sm.add_constant(x)
logit_model=sm.logit(y,x1)
result=logit_model.fit()
print(result.summary2())

Optimization terminated successfully.
    Current function value: 0.604869
    Iterations 5
Results: logit
-----
Model:          logit      No. Iterations:  5.0000
Dependent Variable: coupon      Pseudo R-squared: 0.101
Date: 2019-09-11 12:54 AIC: 126.9739
No. Observations: 100      BIC: 134.7894
Df Model: 2      Log-Likelihood: -60.487
Df Residuals: 97      LL-Null: -67.301
Converged: 1.0000      Scale: 1.0000
-----
            Coef.  Std.Err.      z      P>|z|  [0.025  0.975]
const -2.1464  0.5772 -3.7183  0.0002 -3.2778  1.0150
Card  1.0987  0.4447  2.4707  0.0135  0.2271  1.9703
Spending  0.3416  0.1287  2.6551  0.0079  0.0894  0.5938
-----
```

So this was the output. I am going to explain how we got this G statistics. So look at this value, which I have coloured in the blue colour log likelihood is - 60.487, log likelihood when the variable is not there. That is a log likelihood underscore null is – 67.307.

(Refer Slide Time: 25:05)

G Statistics

$$G = -2 \ln \left[\frac{(\text{likelihood without the variable})}{(\text{likelihood with the variable})} \right].$$

$$G = 2(-60.487 - (-67.301)) = 13.628$$

- The value of G is 13.628, its degrees of freedom are 2, and its p-value is 0.001.
- Thus, at any level of significance $\alpha \geq .001$, we would reject the null hypothesis and conclude that the overall model is significant.

Formula to find out the G statistics is $G = -2\ln$; There is a log likelihood with variable. First one is without variable numerator, the denominator is with variable. So $G = 2$, we got this – 60.487 you see that this value, - 60.487. So when it is null, that value is – 67. 307. So when you find the

difference and multiply by 2, this value is your G value. 13.628. So the value of G is we are getting the same answer. G is 13.628.

It is degrees of freedom for 2, because 2 independent variables and corresponding p-value is, it is 0.001. Thus at any level of significance, since alpha is greater than 0.001 is very low, we would reject null hypothesis and conclude that the overall model is significant. In this class, I have explained when to go for logistic regression equation. When should we go for logistic regression equation, whenever the dependent variable is the categorical variable, we should go for logistic regression equation.

Then I have taken a sample problem. With the help of sample problem, I have used Python to predict the different values that is the various parameters of the logistic regression equation. Then, I have explained what is the G statistics. The G statistics is equivalent to F statistics in our linear regression model. In the linear regression model, F statistics is used to test to the overall significance of the model.

The same way in the logistic regression equation or model, the G statistics is used to test the overall significance, but we have to check the individual significance of each independent variable. That we will continue in the next class. It is equivalent to looking at the t value of our linear regression model. The linear regression model and t statistics is used to test the significance of an individual variable. The same way in our logistic regression equation, the z statistics or the Wald statistics method is used to find out the significance of each independent variable. That we will continue in the next class.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology - Roorkee

Lecture – 39
Logistic Regression - II

In the previous class, we have done the overall significance of the logistic regression model, in this class, we will go for testing the individual significance of all independent variable.

(Refer Slide Time: 00:41)

Chi sq. value of G- Statistic

```
In [5]: import scipy  
         from scipy.stats import chi2  
  
In [7]: chi2.pdf(13.628,2)  
Out[7]: 0.000549145469075383
```

So, the agenda for this class is; testing the significance of logistic regression coefficient then we will do the Python demo on logistic regression. In the previous class I have stopped by saying the G statistics and corresponding p value, that p value has less than 0.05, then we have seen the overall model is significant. So, this is the code from Python to get the for different chi square G value and degrees of freedom.

So, $\text{chi square.pdf}(13.628, 2)$, has our G value in the previous lecture, 2 was our degrees of freedom because there was 2 independent variable, so this was the p value, so the p value is very low, we can say that the model is significant.

(Refer Slide Time: 01:30)

z test- Wald Test

```

In [12]: x = df[['Card','Spending']]
y = df['Coupon']

import statsmodels.api as sm
x1= sm.add_constant(x)
logit_model=sm.logit(y,x1)
result=logit_model.fit()
print(result.summary2())

```

Optimization terminated successfully.
 Current function value: 0.604869
 Iterations 5
 Results: Logit

Model: logit No. Iterations: 5.0000
 Dependent Variable: Coupon Pseudo R-squared: 0.101
 Date: 2019-09-11 12:54 AIC: 126.9794
 No. Observations: 100 BIC: 134.7894
 DF Model: 2 Log-Likelihood: -60.487
 DF Residuals: 97 LL Null: -67.301
 Converged: 1.0000 Scale: 1.0000

	Coef.	Std.Err.	z	P> z	[0.025 0.975]
const	-2.1464	0.5772	-3.7183	0.0002	-3.2778 -1.0150
Card	1.0987	0.4447	<u>2.4707</u>	<u>0.0135</u>	0.2271 1.9703
Spending	0.3416	0.1287	<u>2.6551</u>	<u>0.0079</u>	0.0894 0.5938

Z test or Wald test; z test can be used to determine whether each of the individual independent variable is making significant contribution to the overall model or not. For example, how we got the z value; if you divide 1.0987 divided by 0.447, you will get the z value. Similarly, when you divide 0.3416 by 0.1287, you will get this z value, so corresponding probability you see this one, both the probabilities are less than 0.05. So, we can say both the independent variable is significant, as a whole model also significant, the independent variable in a logistic regression model also significant.

(Refer Slide Time: 02:15)

Strategies

- Suppose Simmons wants to send the promotional catalog only to customers who have a 0.40 or higher probability of using the coupon.
- Customers who have a Simmons credit card:** Send the catalog to every customer who spent \$2000 or more last year.
- Customers who do not have a Simmons credit card:** Send the catalog to every customer who spent \$6000 or more last year.

		Annual Spending						
Credit Card	Yes	\$1000	\$2000	\$3000	\$4000	\$5000	\$6000	\$7000
Yes	No	0.3305	0.6039	0.6043	0.5793	0.5594	0.7315	0.7301
No	No	0.1413	0.1860	0.2657	0.3144	0.3922	0.4759	0.5610

Once we came to know both are significant, then we will go for interpretation of its output, so what kind of different strategies that company has to adopt, so that they can improve their revenue by selling more coupons. Suppose Simmons wants to send the promotional catalog

only to customers who have a 0.40 or higher probability of using the coupon. So, what will happen, you look at where this 0.4, here this one 0.4, those who are having credit card.

Those who are not having credit card, 0.4 is here, so what interpretation from this table is; whoever having the credit card and whose spending is 2,000 dollar and above, for them you can send the coupon, they will use the coupon. Those who are not having the coupon but their spending is above 6,000 dollar, for them you can send the coupon, so that they will use that one.

So, this is the managerial interpretation, so customers who have a Simmons credit card, send the catalog to every customers who spend 2,000 dollar or more last year because the 0.4 is the cut-off, customers who do not have the Simmons credit card send the catalog to every customer who spent 6,000 dollar or more in the last year, so that is the strategy for the promotion.

(Refer Slide Time: 03:49)

Interpreting the Logistic Regression Equation

$$\text{odds} = \frac{P(y = 1 | x_1, x_2, \dots, x_p)}{P(y = 0 | x_1, x_2, \dots, x_p)} = \frac{P(y = 1 | x_1, x_2, \dots, x_p)}{1 - P(y = 1 | x_1, x_2, \dots, x_p)}$$

$$\frac{P}{1-P}$$

Now, we will go for interpreting the logistic regression equation; for that there is a close connection between odd. What is odd is, probability of success divided by probability of not getting success, so generally the odd is P divided by $1 - P$ that is odd, so probability of y equal to 1 is the success, probability of y equal to 0 is not success. So, probability of y equal to 1 given that a different independent variable that is a numerator P , so $1 - P$ of y equal to 1 that is $1 - P$, this function is called odds function.

(Refer Slide Time: 04:33)

Odd ratio

$$\text{Odds Ratio} = \frac{\text{odds}_1}{\text{odds}_0}$$

- The **odds ratio** measures the **impact on the odds of a one-unit increase** in only one of the independent variables.

Then from odds function, we are going to the next term odds ratio because this odds ratio is very, very helpful to explain the coefficient of logistic regression equation. So, the odds ratio is odds 1 divided by odds 0, that means when 0th level what is the odd, when the first level what is the odd, so 0th level means for example, those who are not having credit card that is odds is 0.

Those who are having the credit card suppose, we are jumping from one level to another level, odds 1 is those who are having the credit card, so the odd ratio measures the impact of odds of 1 unit increase in only one of the independent variable, so this odds ratio will help us to interpret to the logistic regression equation, if the independent variable is increased by 1 unit, what is the effect of that on the dependent variable will interpret this.

(Refer Slide Time: 05:33)

Interpretation

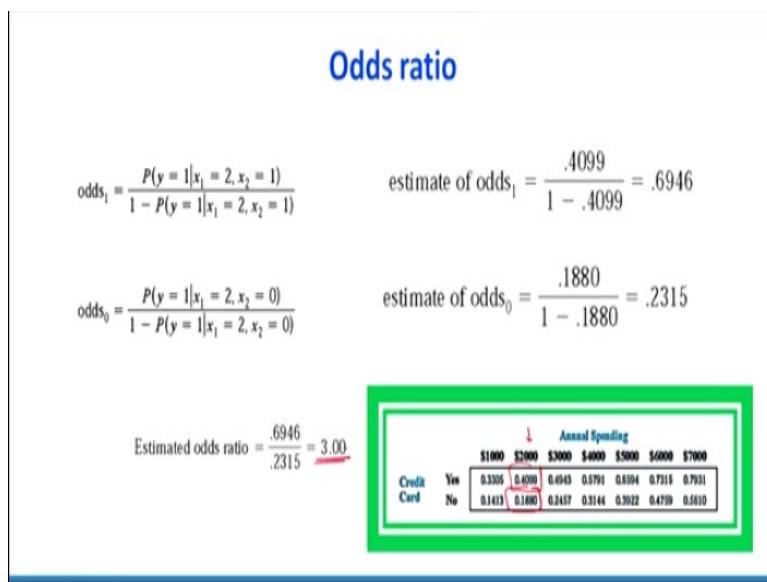
- For example, suppose we want to compare the odds of using the coupon for customers who spend \$2000 annually and have a Simmons credit card ($x_1=2$ and $x_2=1$) to the odds of using the coupon for customers who spend \$2000 annually and do not have a Simmons credit card ($x_1=2$ and $x_2=0$).
- We are interested in interpreting the effect of a one-unit increase in the independent variable x_2 .

1 → Credit card
0 no credit

What is interpretation? For example, suppose we want to compare the odds of using the coupon for customers who spent 2,000 dollar annually and have a Simmons credit card, so that is this category; x_1 is a 2000 having the credit card to the odds of the coupon for customers who spent 2,000 dollar annually and do not have the Simmons credit card. So, what will happen here; x_2 will becomes 0, so x_1 equal to 2, x_2 equal to 0, so that is the second category is odd 0, the first one is odds 1.

So, the ratio of that 2 is called odd ratio, so we are interested in interpreting the effect of 1 unit increase on the independent variable x_2 , so 1 unit increase means here person having not credit card to a person having the credit card, so the 0 level to 1 level. The 0 level is no credit card, the 1 level is credit card, okay. So, what is the impact of this one on the estimated value of y ?

(Refer Slide Time: 06:48)



First, we will see odds 1; odds 1 is a person having the correct card, so P of y equal to 1, x_1 equal to 2, spending is 2,000 dollar, x_2 equal to 1 having the credit card, this is the numerator, probability of success we can say probability of not success, $1 - P$ of y equal to 1, x_1 equal to 2, x_2 equal to 1. So, what will happen; you can substitute this, when x_1 equal to 2, this value and having the credit card that is this 0.0499.

So when you substitute this 0.4099 divided by $1 - 0.4$, we are getting 0.6946, we will go to the odds 0; 0th level. So, P of y equal to 1, x_1 equal to 2, x_2 equal to 0, this case similar to previous one spending amount is same but he is not having credit card, for that the probability is numerator is P , denominator is $1 - P$. So, what is that category; this one,

0.1880, so 0.1880 divided by 1 – 0.1880 that is giving 0.2315, so when you divide this 0.6946 divided by 0.2315, this 3 so, the 3 is very useful for interpreting.

(Refer Slide Time: 08:21)

Odds ratio – Interpretation

- The estimated odds in favor of using the coupon for customers who spent \$2000 last year and have a Simmons credit card are 3 times greater than the estimated odds in favor of using the coupon for customers who spent \$2000 last year and do not have a Simmons credit card.

What is the meaning of 3 is; the estimated odds in favour of using the coupon for customers who spent 2,000 dollar last year and have a Simmons credit card are 3 times greater than the estimated odds in favour of using the coupon for customers who spent to 2,000 dollar last year and do not have the Simmons credit card. So that means, expenditure is, the spending amount is same. But when you sent this coupon to the person who is having the credit card, there is a 3 times more chance that person will use that coupon, okay that is the meaning of this odds ratio.

(Refer Slide Time: 09:03)

Odds ratio – Interpretation

- The odds ratio for each independent variable is computed while holding all the other independent variables constant.
- But it does not matter what constant values are used for the other independent variables.
- For instance, if we computed the odds ratio for the Simmons credit card variable (x_2) using \$3000, instead of \$2000, as the value for the annual spending variable (x_1), we would still obtain the same value for the estimated odds ratio (3.00).
- Thus, we can conclude that the estimated odds of using the coupon for customers who have a Simmons credit card are 3 times greater than the estimated odds of using the coupon for customers who do not have a Simmons credit card.

The odds ratio for each independent variable is computed while holding all other independent variable is constant for example, in the previous case also where the expenditure; the amount spent that is expenditure is taken as the constant, we have interpret only for a person having the credit card or not having the credit card, it does not matter constant values are used for other independent variables.

So, we do not bother about the constant variables for instance, if we computed the odd ratio for Simmons credit card variable x_2 instead of 2,000 you say, 3,000 dollar expenditure, instead of 2,000 as the value of the annual spending variable is x_1 , we would still obtain the same value of estimated odd ratio, so the constant does not matter.

Thus we can conclude that the estimated odds of using coupon for customers who have a Simmons credit card are 3 times greater that is important interpretation, 3 times greater than the estimated odds of using the coupon for customers who do not have the Simmons credit card. So, you have to target to a person who has the credit card when you target to them, there is a 3 times more chance that people will use when compared to those who are not having credit card they will use the coupon.

(Refer Slide Time: 10:27)

Relationship between the odds ratio and the coefficients of the independent variables

		Annual Spending						
		\$1000	\$2000	\$3000	\$4000	\$5000	\$6000	\$7000
Credit Card	Yes	0.3305	0.4599	0.6943	0.8791	0.8994	0.7315	0.7931
	No	0.1413	0.1880	0.2457	0.3144	0.3922	0.4759	0.5610

Odds ratio = e^{β_i}

Estimated odds ratio = $e^{\beta_1} = e^{341643} = 1.41$

Estimated odds ratio for x_2 is

Estimated odds ratio = $e^{\beta_1} = e^{1.09873} = 3.00$

Now, another very useful relationship instead of finding odd ratio that way, there is a connection between the coefficient of logistic regression equation and the odds ratio, that is this one; e to the power beta i, what is the title says; relationship between odds ratio and the coefficient of independent variable, the beta i is called the coefficient of independent variable.

So, if you want to know the estimated odds ratio for x_1 variable that is the amount spent, e to the power b_1 , in our equations b_1 is 0.34, where we got this one b_1 , I am going back this one, so here we are taking x_1 variable, this is x_2 variable, so e to the power 0.3416 that will give you the odd ratio for this variable because in some software packages for example, Minitab they directly give the odd ratio for each independent variable.

But in Python we can calculate the odd ratio using that relationship e to the power β_1 for example, if we want to know odd ratio for card, those who are having card or not, so e to the power 1.0987 that will become 3, I will show you that one yeah, see that b_2 ; e to the power b_2 , e to the power 1.09873 is 3, so odd ratio we can directly get from the coefficient of logistic regression equation.

(Refer Slide Time: 12:05)

Effect of a change of more than one unit in Odd Ratio

$$x_1 = 2 \quad x_1 = 3 \quad x_1 = 0 \quad x_1 = 1$$

- The odds ratio for an independent variable represents the change in the odds for a one unit change in the independent variable holding all the other independent variables constant.
- Suppose that we want to consider the effect of a change of more than one unit, say c units.
- For instance, suppose in the Simmons example that we want to compare the odds of using the coupon for customers who spend \$5000 annually ($x_1 = 5$) to the odds of using the coupon for customers who spend \$2000 annually ($x_1 = 2$).
- In this case $c = 5 - 2 = 3$ and the corresponding estimated odds ratio is

Now, so far we have seen there is 1 unit is change then we have seen, what is the corresponding effect on the dependent variable, sometime what will happen; what is the meaning of 1 unit change means, suppose we have taken for x_2 equal to 0 and 1, we have seen 1 unit change, if there is 1 unit change, we have seen effect of that on the dependent variable.

For example, this is a discrete variable, if the independent variable is a continuous variable for example, say x_1 is amount spent suppose, somebody is spending 2,000 dollar, what is the probability that people will use the coupon, so we can see 2,000 to 3,000 that is x_1 equal to 2 to 3, if it instead of 1 unit jump, we can go for 6 unit or 5 unit jump at a time and

corresponding interpretation we can find out that is the meaning of that change of more than 1 unit in the odd ratio.

The odd ratio for an independent variable represents the change in odds of 1 unit each change in the independent variable holding all other independent variables are constant. Suppose, we want to consider the effect of a change of more than 1 unit for example, c unit instead of 2 to 3, I want to say 2 to 5 for instance, suppose the Simmons example that we want to compare the odds of using the coupon for customers who spent 5,000 dollar annually to the odds of using the coupon for customers who spent 2,000, the increment is not 1. Because it is a 3 okay, in this case c equal to 5 - 2 is 3 and the corresponding estimated odd ratio is very, very useful.

(Refer Slide Time: 14:05)

Effect of a change of more than one unit in Odd Ratio

$$e^{cb_1} = e^{3(341643)} = e^{1.0249} = \underline{2.79}$$

- This result indicates that the estimated odds of using the coupon for customers who spend \$5000 annually is 2.79 times greater than the estimated odds of using the coupon for customers who spend \$2000 annually.
- In other words, the estimated odds ratio for an increase of \$3000 in annual spending is 2.79

So, e to the power c , this c is how much is we are increasing, so when you multiply by 3 of this one, we are getting the odd ratio is 2.79, this result indicates that the estimated odds of using the coupons for the customers who spend 5,000 dollar annually is 2.79 times greater than the estimated odds of using the coupon for customers who spend only 2,000 dollar annually.

You see that here the increment is not the unit increment, it is the 3 times increment in other words, the estimated odd ratio for increase of 3,000 in annual spending yeah, it is a 3 unit means, 3000 is 2.79.

(Refer Slide Time: 14:49)

Logit Transformation

- An interesting relationship can be observed between the odds in favor of $y = 1$ and the exponent for 'e' in the logistic regression equation

$$\ln(\text{odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- This equation shows that the natural logarithm of the odds in favor of $y = 1$ is a linear function of the independent variables.
- This linear function is called the **logit** $\rightarrow g(x_1, x_2, \dots, x_p)$ to denote the logit.

Then we will come to some theory portions of this logistic regression equation, first we will see what is the logit transformation. An interesting relationship can be observed between odds in favour of y equal to 1 and the exponent of e in the logistic regression equation, so when you say, as I told you previously probability of success, probability of non-success when you take log of that, that is nothing but a logit function.

So, log of odds equal to beta 0 + beta 1 x_1 + beta 2 x_2 + beta p x_p , this equation shows that the natural logarithm of the odds in favour of y equal to 1 is a linear function of independent variable. So, why we are taking log, so that will become a linear function, this linear function is called logit generally, in the custom is g of x_1, x_2 up to x_p to denote the logit function. So, when you take logit function it will become linear, so interpretation is easy.

(Refer Slide Time: 16:04)

Estimated Logit Regression Equation

$$g(x_1, x_2, \dots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$E(y) = \frac{e^{g(x_1, x_2, \dots, x_p)}}{1 + e^{g(x_1, x_2, \dots, x_p)}}$$

$$\hat{y} = \frac{e^{(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}} = \frac{e^{\hat{g}(x_1, x_2, \dots, x_p)}}{1 + e^{\hat{g}(x_1, x_2, \dots, x_p)}}$$

So, how to estimate the logistic regression equation, we know this a logit function, if you want to know the expected value of this logistic function is nothing but e to the power g of x_1, x_2 up to x_p divided by $1 + e$ power of g of x_1, x_2 up to x_p , so you can expand this with the help of sample data; b_0, b_1, b_2 up to b_p , so this is the your sample value, with the help of the sample data we can predict the population parameter.

Sample; generally, the name parameter is used only for the population not for the sample, so e to the power; when I write the hat symbol it is the estimated value, y hat, g hat, so you can this can be written as because this is right, this can be is nothing but this value, so we can do logit function, so e to the power g hat x_1, x_2 up to x_p divided by $1 + e$ to the power g hat this one.

(Refer Slide Time: 17:15)

$$\hat{g}(x_1, x_2) = -2.14637 + 0.341643x_1 + 1.09873x_2$$

$$\hat{y} = \frac{e^{\hat{g}(x_1, x_2)}}{1 + e^{\hat{g}(x_1, x_2)}} = \frac{e^{-2.14637 + 0.341643x_1 + 1.09873x_2}}{1 + e^{-2.14637 + 0.341643x_1 + 1.09873x_2}}$$

Why we are taking e to the power and taking log; to make it linear, in our problem so far which we have discussed, so this was our logit equation is $-2.14, 0.34164 x_1 + 1.09873 x_2$, and if you want to predict y , here y where is the probability value, e to the power -2141 , this we got this answer.

(Refer Slide Time: 17:40)

G vs Z

- Because of the unique relationship between the estimated coefficients in the model and the corresponding odds ratios, the overall test for significance based upon the G statistic is also a test of overall significance for the odds ratios.
- In addition, the z test for the individual significance of a model parameter also provides a statistical test of significance for the corresponding odds ratio.

$$\begin{array}{l} G - F \\ Z - t \end{array}$$

Very important things; we will compare what is the purpose of G statistics and Z statistic, as I told you because of the unique relationship between estimated coefficient in the model and the corresponding odds ratio, the overall test is very important; the overall test for the significance based upon G statistics also is test of a overall significance for the odd ratio but the z test or the Wald test for the individual significance of model parameters also provide a statistical test of significance for the corresponding odd ratio.

This is similar to G is similar to F test, z is similar to t test in the; this is for, the right side one is for linear regression, the left side one is for logistic regression. Now, we will go for Python the data which I have explained to you which I have brought you in the screenshot, I will run that model then I will show you how to do the logistic regression using Python.

(Refer Slide Time: 18:55)

The screenshot shows a Jupyter Notebook interface with the following code:

```
In [1]: import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
from sklearn import linear_model  
import statsmodels.api as sm  
from sklearn.metrics import mean_squared_error  
  
In [2]: df = pd.read_excel('Simmons.xls')  
  
In [3]: df  
  
In [4]: plt.scatter(df.Spending, df.Coupon, marker='*', color='red')  
  
In [5]: x = df[['Card', 'Spending']]  
y = df['Coupon']  
  
import statsmodels.api as sm  
x1= sm.add_constant(x)  
logit_model=sm.Logit(y,x1)  
result=logit_model.fit()  
print(result.summary2())
```

Now, we will go to our Python environment, then I will teach you how to do the logistic regression, so what are the libraries required? You need pandas, you need numpy, you need matplotlib.pyplot, you need sklearn for doing linear model, you can import statsmodels.api, then from sklearn.metrics import mean_squared_error, the file name is Simmons.xls, as I told you this was taken from Anderson, Sweeney and Williams book.

(Refer Slide Time: 19:26)

```

import matplotlib.pyplot as plt
from sklearn import linear_model
import statsmodels.api as sm
from sklearn.metrics import mean_squared_error

In [2]: df = pd.read_excel('Simmons.xls')

In [3]: df
Out[3]:
   Customer  Spending  Card  Coupon
0         1      2.291    1      0
1         2      3.215    1      0
2         3      2.135    1      0
3         4      3.924    0      0
4         5      2.528    1      0
5         6      2.473    0      1
6         7      2.384    0      0
7         8      7.076    0      0
8         9      1.182    1      1
9        10      3.345    0      0

```

So, this was the data, so what is happening I am scrolling, there is a 100 data set is there, so what are the variable is there; customer number is there, 1, 2, 3 up to 100, spending; how much they spend last time, then whether the possession of the card or not, if 1 means they are having the card, 0 means not having the card, then coupon; whether they have used the coupon, 0 means not use the coupon, 1 means uses the coupon.

(Refer Slide Time: 19:56)

```

plt.scatter(df['Spending'], df['Coupon'], marker='*', color='red')

In [4]: x = df[['Card', 'Spending']]
y = df['Coupon']

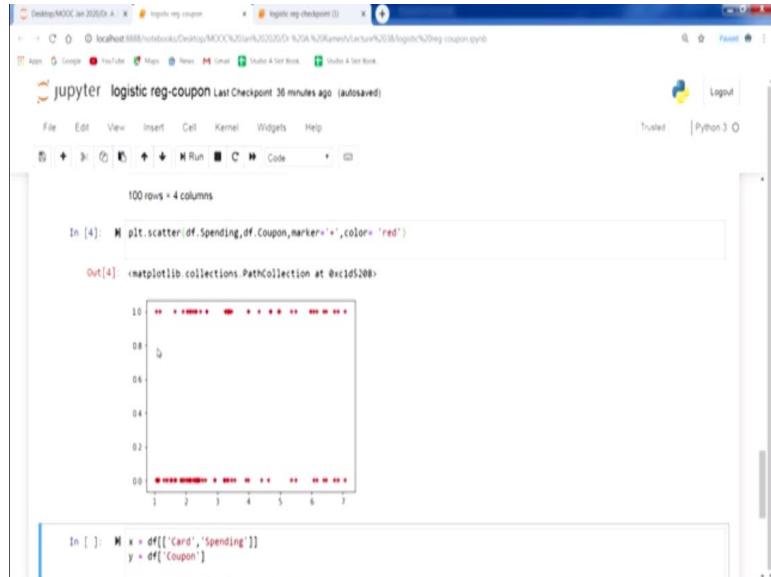
import statsmodels.api as sm
x1 = sm.add_constant(x)
logit_model = sm.Logit(y, x1)
result = logit_model.fit()
print(result.summary2())

In [5]: import scipy
from scipy import stats

```

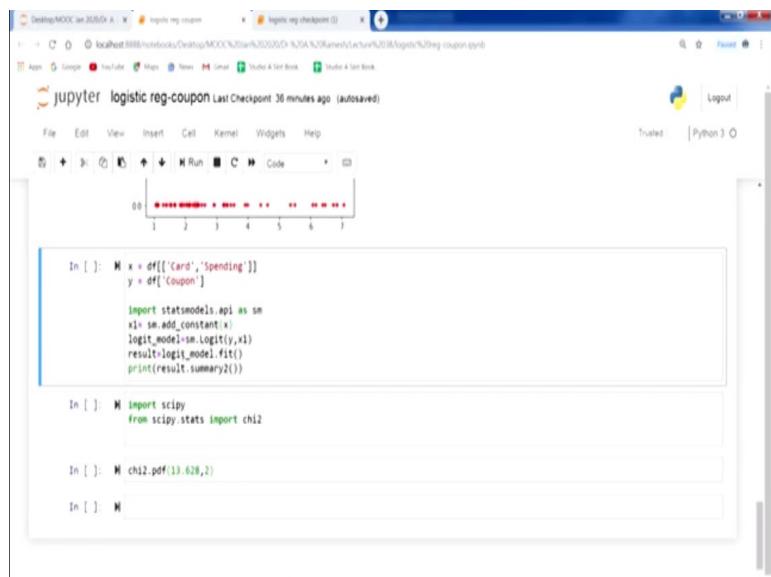
First, we will do the scatterplot between spending and coupon, when you look at this data spending is a continuous variable, coupon is the categorical variable, coupon is our dependent variable.

(Refer Slide Time: 20:13)



So, when you run this you see that you are getting this way, so for this kind of model whether there is a 2 possibility, it is 1 or 0, in between there is no possibility, so the linear regression model is not valid here, we should go for logistic regression model that is one point. The another point is here the assumption is a linear regression model the error term will follow normal distribution but the logistic regression model error term will follow binomial distribution, so you cannot go for linear regression model.

(Refer Slide Time: 20:47)



In x, I am taken card and spending, the Y is a dependent variable is coupon, so x1 equal to sm.add_constant(x), logit_model = sm.logit(y, x1), result = logit_model.fit(), so I am going to summary of the logistic regression.

(Refer Slide Time: 21:10)

```

Model: Logit Pseudo R-squared: 0.101
Dependent Variable: Coupon AIC: 126.9739
Date: 2019-09-12 18:28:01 BIC: 134.7894
No. Observations: 100 Log-Likelihood: -60.487
DF Model: 2 LL-Null: -67.301
DF Residuals: 97 LLR p-value: 0.0010981
Converged: 1.0000 Scale: 1.0000
No. Iterations: 5.0000
*****
```

	Coeff.	Std.Err.	z	P> z
const	-2.1464	0.5772	-3.7183	0.0002
Card	1.0987	0.4447	2.4707	0.0135
Spending	0.3416	0.1287	2.6551	0.0079
				0.0894 0.5938

```

In [ ]: %import scipy
         from scipy.stats import chisq
```

```

In [ ]: chisq.pdf(13.628,2)
```

```

In [ ]:
```

So, summary of logistic regression is see the constant value we need not bother about this, for card it is 1.0987, for spending it is 0.3416, so after getting this output what do you have to see; you have to check the overall significance with the help of G statistics that it is not here but you have to find out how? 2 multiplied by - 60.487 minus, - 67.301, so that value is your G value.

For that G value you have to find out by having 2 degrees of freedom, that G value is nothing but your chi square value, you have to find out what is the corresponding p value, with the p value is less than 0.05, we can say the overall model is significant. The next aspect is checking whether each independent variable is significant or not that is done with the help of Wald test.

So, you can look at here, it is the p value is 0.01 less than 0.05, for second variable also the p value is less than 0.05, then we can say both the variable is significant. Suppose, if you want to interpret this model with the help of odd ratio, what you have to do; when you take e to the power this beta 1 that is e to the power 1.0987, you will get a corresponding odd ratio that is used to explain 1 unit increase.

Suppose, the card person is not having card to having the card, so what was the corresponding effect on the dependent variable that can be found out. The spending is a continuous variable, here also we can find out e to the power -0.3416 will give you the odd ratio, so that will help you to interpret, suppose a person is spending 2,000, another person is spending 3,000.

Suppose, there is 1 unit jump what is the corresponding chances because of that 1 unit jump that the person will use the coupon suppose, if there is a c unit jump simply you have to see; you have to find out e to the power c of that is a c of 0.3416, so you will get a c unit odd ratio that you can directly interpret it.

(Refer Slide Time: 23:33)

The screenshot shows a Jupyter Notebook interface with two code cells and their outputs.

Cell 1 Output:

```
DF Residuals: 97 LLR p-value: 0.0010981
Converged: 1.0000 Scale: 1.0000
No. Iterations: 5.0000
*****
      Coef. Std Err. z P>|z| [0.025 0.975]
*****
const -2.1464 0.5772 -3.7183 0.0002 -3.2778 -1.4150
Card  1.0987 0.4447 2.4707 0.0135 0.2271 1.9703
Spending 0.3416 0.1287 2.6551 0.0079 0.0894 0.5938
*****
```

Cell 2 Output:

```
In [6]: M import scipy
         from scipy.stats import chi2

In [7]: M chi2.pdf(13.628,2)
Out[7]: 0.000549145469075183
```

This is the code to check the G value for example, as I told you previously the G value is 13.628, it is mentioned in my PPT also, so chi square value is 13.628 and the degrees of freedom is 2, why it is 2 because there are 2 independent variable, so corresponding p value is 0.0054 that is less than 0.05, overall model is significant. In this class I have explained how to test the significance of each independent variable in a logistic regression equation that I have done with the help of z statistics otherwise, Wald statistics.

Then I have explained what is the odds, then I have explained what is the odds ratio, then I explained how to use this odds ratio to interpret the coefficient of logistic regression equation, at the end I have used Python and I have shown how to use Python for running logistic regression equation. Thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology - Roorkee

Lecture – 40
Linear Regressions Model VS Logistic Regression Model

In this class, we are going to compare logistic regression versus linear regression because it is very important to understand how this linear regression and logistic regressions are connected. If you understand the relationship between this linear and logistic regression, it is easy to interpret the meaning of logistic regression. So, agenda of this class is comparison of linear regression model and logistic regression model.

(Refer Slide Time: 00:56)

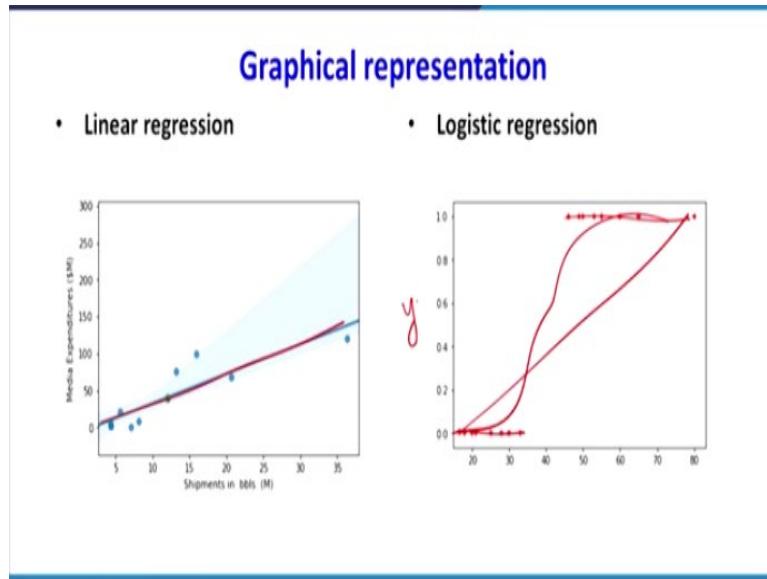
Estimating the relationship	
Linear regression model	Logistic regression model
<ul style="list-style-type: none">• $Y_1 = X_1 + X_2 + \dots + X_n$• Where ,<ul style="list-style-type: none">– Y_1 = continuous data– Independent variables = nonmetric and metric	<ul style="list-style-type: none">• $Y_1 = X_1 + X_2 + \dots + X_n$• Where ,<ul style="list-style-type: none">– Y_1 = Binary nonmetric– Independent variables = nonmetric and metric

We will see the first relationship, first difference, estimating the relationship, when you look at the linear regression model, we used to write Y_1 equal to $X_1 + X_2 + \dots + X_n$, where Y_1 is a continuous data that is dependent variable, X_1, X_2 are independent variable, this independent variable it can be continuous we can call it as metric, otherwise it may be a discrete, we can call it as non-metric.

The linear regression model, if the independent variable is discrete variable, we can use the concept of dummy variable regression, whereas in logistic regression model, the general model is Y_1 equal to $X_1 + X_2 + \dots + X_n$, where the Y_1 is a binary variable, it is a nonmetric binary

variable. The independent variable can be continuous or discrete, we can call it other way; it may be a metric or nonmetric that is a basic difference between linear regression model and logistic regression model.

(Refer Slide Time: 02:02)



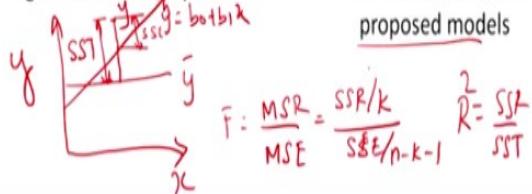
The another difference is if you plot a simple independent variable and the dependent variable for a linear regression, you may able to connect all the points this way but in logistic regression, value of Y can be only 2 possibilities may be 0 or 1, you may get this kind of relationship. What is a meaning here is you cannot form a linear relationship this way, you have to form a; a kind of an S shaped curve that is another difference between linear and logistic regression. What is that in the y axis, see 0 to 300 for linear for example, here you see that it is a possibility, not only that the y value is nothing but the probability but here their y value is the actual values.

(Refer Slide Time: 02:53)

Correspondence of Primary Elements of Model Fit

Linear Regression

- Total sum of squares SST
- Error sum of squares SSE
- F test of model fit
- Coefficient of determination (R^2)
- Regression sum of squares SSR



Logistic Regression

- -2LL of base model
- -2LL of proposed model
- Chi-square test of -2LL difference (6)
- Pseudo R^2 measures
- Difference of -2LL for base and proposed models

$$F: \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)} \quad R^2 = \frac{SSR}{SST}$$

Correspondence of primary elements of model fit between linear regressions, logistic regression, we used to write SST; total sum of square that is for linear regression, the equivalent term for logistic regression is -2LL that is the log likelihood of base model. Here, we have to write SSE; error sum of square, the equivalent term in the logistic regression is -2LL of proposed model. I have explained, what is the meaning of log likelihood in my previous lectures.

In this lecture also, I will show you the software output where we can get it this log likelihood value. We know that in a simple linear regression, the meaning of SST is like this, say this is y bar, this is y , suppose a line goes like this, this is our predicted value $b_0 + b_1 x$, this is our x value, this is our y value. So, this distance was our SST, the equivalent value in the logistic regression is -2LL.

Similarly, we have seen SSE, this unexplained this length, this length in the regression equation is SSE; error sum of square are unexplained variance portions, in the linear regression model to test the overall fit; model fit, we have used F test. There what was the F test in the linear regression model, F is MSR divided by MSE, what is MSR; MSR is SSR divided by k; number of independent variable, divided by SSE $n - k - 1$.

Sometimes, some books they use k, some books they use p to explain the number of independent variables that is a F value. The equivalent test in logistic regression is chi square test of -2LL

difference that value is nothing but your G, in the linear regression to explain the model fit, the goodness of the model the term used is coefficient of determination, R square. So, what is R square?

R square is SSR divided by SST, regression sum of square divided by total sum of square, otherwise explained variance divided by total variance. The equivalent term in logistic regression is pseudo R square; I will explain what is the formula for finding pseudo R square in coming slides. Here, we use SSR; SSR is regression sum of square, the equivalent term for logistic regression is difference of – 2LL for base and proposed model.

I will explain the meaning of base and proposed model, base means when there is no independent variable corresponding log likelihood values called base value, when you introduce any 1 independent variable, after introduction of independent variable, the corresponding log likelihood values called model; model log likelihood, I will explain this detail in coming slides.

(Refer Slide Time: 06:54)

Objective of logistic regression

- Logistic regression is identical to discriminant analysis in terms of the basic objectives it can address
- Logistic regression is best suited to address two research objectives:
 - Identifying the independent variables that impact group membership in the dependent variable
 - Establishing a classification system based on the logistic model for determining group membership

Then with respect to objective of logistic regression, how the linear and logistic regression differs. Logistic regression is identical to discriminant analysis in terms of basic objectives it can address. There is a one technique called discriminant analysis, the basic difference is in logistic regression, we had only 2 levels; 0 or 1 but in the discriminant analysis, we can have more than 1 level that case is called discriminant analysis.

This I did not cover it but this is the concept behind of discriminant analysis, so logistic regression is identical to discriminant analysis in terms of basic objective it can address; still we go for logistic regression. If there are 2 category, we can go for discriminant analysis also but still we prefer logistic regression because it is best suited to address 2 research objectives, one is identifying the independent variables that impact group membership in the dependent variable. Another one is establishing a classification system based on the logistic model for determining group membership.

(Refer Slide Time: 08:18)

The fundamental difference

- Logistic regression differs from linear regression, in being specifically designed to predict the probability of an event occurring (ie., the probability of an observation being in the group coded 1)
- Although probability values are metric measures, there are fundamental differences between linear regression and logistic regression



There are some more reason is there, the fundamental difference between logistic and linear regression is; logistic regression differs from linear regression in being specifically designed to predict the probability of an event occurring, so the y value is nothing but the probability that is the probability of observation being group coded 1 or not, although the probability values are metric measures, there are fundamental differences between linear and logistic regression. Even though, we can say the probability value is metric, so there is a 2 possibility, it may be 0 or 1, so we may get different probability, when we go for logistic regression.

(Refer Slide Time: 09:02)

Log likelihood

- Measure used in logistic regression to represent the lack of predictive fit
- Even though this method does not use the least squares procedure in model estimation, as is done in linear regression, the likelihood value is similar to the sum of squared error in regression analysis

SSE

Then, log likelihood; measures used to logistic regression to represent lack of predictive fit, so the log likelihood is used to measure how much lack of fit is there, even though this method does not use in least square procedure in model estimation as is done in linear regression, the likelihood value is similar to sum of squared error. If the log likelihood value is lesser, it is better because in the regression, we try to have sum of squared error SSE lesser it is better. Similar to that in logistic regression, if you are getting smaller value of log likelihood it is better that is a good model.

(Refer Slide Time: 09:52)

Logistic vs discriminant

- Logistic regression may be preferred for two reasons
- First, discriminant analysis relies on strictly meeting the assumptions of
 - Multivariate normality and equal variance
 - Covariance matrices across groups
 - Assumptions that are not met in many situations
- Logistic regression does not face these strict assumptions and is much more robust when these assumptions are not met, making its application appropriate in many situations

Now, we will compare when should we go for logistic regression, when should we go for discriminant analysis, even 2 slides before also, I have explained comparison between logistic

and discriminant analysis. Discriminant analysis we can have more than 2 levels, 3 levels or 4 levels. The problem related to logistic regression can be solved with the help of discriminant analysis, where there are 2 levels.

We can say logistic regression is a special case of discriminant analysis but we would not go for discriminant analysis, we will go for logistic regression, there are some reason is there. See that the logistic regression may be preferred for 2 reasons. First; discriminant analysis relies on strictly meeting the assumption of multivariate normality and equal variance that is the first assumption for the discriminant analysis.

That means, the data has to follow normality and it has to have equal variance and the covariance matrices across groups is necessary, when we go for discriminant analysis. Assumptions that are not met in many situations, a real time problems we cannot have this assumptions, in that situation whenever there is a 2 level in the dependent variable instead of going for discriminant analysis, we can go for logistic regression.

Because these assumptions are not required for logistic regression otherwise, we can say logistic regression is more robust than discriminant analysis when there is an only 2 category in the dependent variable. The next point; the logistic regression does not face these strict assumption and is much more robust when these assumptions are not met, making its application appropriate in many situations, that is why we are going for logistic regression over discriminant analysis.

(Refer Slide Time: 11:56)

Logistic vs discriminant

- Second, even if the assumptions are met, many researchers prefer logistic regression because it is similar to multiple regression
- It has straightforward statistical tests, similar approaches to incorporating metric and nonmetric variables and nonlinear effects, and a wide range of diagnostics
- Logistic regression is equivalent to two-group discriminant analysis and may be more suitable in many situations

Another point is even though, the assumptions are met, many researchers prefer logistic regression because it is similar to multiple regression, many possibility to interpret the result, it has straightforward statistical test, similar approaches to incorporating metric and nonmetric variables and nonlinear effects and a wide range of diagnostics. Logistic regression is equivalent to 2 groups; this point which I am trying to say, 2 group discriminant analysis may be more suitable in many situations.

(Refer Slide Time: 12:38)

Logistic vs discriminant : Sample size

- One factor that distinguishes logistic regression from the other techniques is its use of maximum likelihood (MLE) as the estimation technique
- MLE requires larger samples such that, all things being equal, logistic regression will require a larger sample size than multiple regression
- As for discriminant analysis, there are considerations on the minimum group size as well

With respect to sample size which is better; logistic or discriminant analysis, one factor that distinguishes logistic regression from the other techniques is its use of maximum likelihood as the estimation technique. Maximum likelihood estimation requires larger sample such that all

things being equal, logistic regression will require a larger sample size than multiple regression, as of discriminant analysis there are considerations on minimum group size as well.

The another when we go for logistic regression, one point is that you need to have large sample size because it follow maximum likelihood estimate because the value of maximum likelihood estimate is sensitive to the sample size or degrees of freedom.

(Refer Slide Time: 13:30)

Logistic vs discriminant : Sample size

- The recommended sample size for each group is at least 10 observations per estimated parameter 3 -
- This is much greater than multiple regression, which had a minimum of five observations per parameter, and that was for the overall sample, not the sample size for each group, as seen with logistic regression

Regression: 1-5
L.R.: 1:10

The recommended sample size for each group is at least 10 observations per estimated parameter, when we go for logistic regression that means, if you are capturing 1 variable, you need to have 20 observations. If you are capturing 3 variables, you have to have 30 observations, this is the thumb rule; this is much greater than multiple regressions which had minimum 5 observations per parameter.

That was for the overall sample not the sample size of each group as seen in the logistic regression, so what the point here is if it is multiple regression for 1 variable, you need to have, you can have 5 respondent rated this regression when you go for logistic regression in to help for one variable in the top 10 respondent where it is regression. When you go for logistic regression, you need to have, for 1 variable, you need to have 10 respondent.

(Refer Slide Time: 14:32)

Determination of coefficients

Linear regression

- R^2
- $r^2 = \text{SSR}/\text{SST}$

where:

SSR = sum of squares due to regression

SST = total sum of squares

Logistic regression

$$R^2_{\text{Logit}} = \frac{-2LL_{\text{null}} - (-2LL_{\text{model}})}{-2LL_{\text{null}}}$$

Where:

LL = Loglikelihood

$-2LL_{\text{null}}$ = -2LL of base model

$-2LL_{\text{model}}$ = -2LL of proposed model

Then the goodness of fit of the both linear and logistic regression, as I told you the linear regression, the R square that is a coefficient of determination is measured by SSR divided by SST, regression sum of square divided by total sum of square. The equivalent term in logistic regression is pseudo R square, otherwise R square logit equal to $-2LL$ for null model, $-2 \log \text{likelihood}$ minus of $-2LL$ log likelihood for model divided by $-2 \log \text{likelihood}$ for null.

This value I will show what is the null LL and model LL, so LL is likelihood, if I say $-2LL$ of log likelihood of base model without any independent variable, $-2LL$ model is meaning here is for model means, it is a proposed model, when you bring a new dependent variable into the logistic regression model that time what was the corresponding log likelihood value that is called this model value for log likelihood.

(Refer Slide Time: 15:54)

Determination of coefficients

- Linear regression

OLS Regression Results						
Dep. Variable:	Media Expenditure (M)					
R squared:	0.783					
Model:	OLS					
Method:	Least Squares					
F statistic:	28.93					
Date:	Wed, 10 Oct 2018					
Prob (F-statistic):	0.00003					
Time:	10:16:26					
Log Likelihood:	-44.355					
No. Observations:	10					
AIC:	92.71					
DF Residuals:	8					
BIC:	93.32					
DF Model:	1					
Covariance Type:	oprobust					
coef std err t P> t [0.025 0.975]						
Intercept	7.6277	11.405	0.664	0.529	-34.112	10.857
Shippments in Mdn (M)	4.0065	0.745	5.378	0.001	2.209	5.724
Omnibus: 4.361 Durbin-Watson: 1.473						
Prob(Omnibus): 0.113 Jarque-Bera (JB): 2.129						
Skew: 1.129 Prob(JB): 0.345						
Kurtosis: 2.925 Cond. No.: 24.6						

- Logistic regression

Results: Logit						
Model:	Logit					
Pseudo R-squared:	0.192					
Dependent Variable:	Coupon					
AIC:	12.8004					
Date:	2019-09-08 11:07:41					
IC:	12.6916					
No. Observations:	10					
Log likelihood:	-4.8432					
Df Model:	1					
L-L null:	-5.0000					
Df Residuals:	8					
LR p-value:	0.16560					
Converged:	1.0000					
Scale:	1.0000					
No. Iterations:	7.0000					
Coef. Std.Err. z P> z [0.025 0.975]						
Spending	-0.6318	0.4566	-1.3838	0.1664	-1.5267	0.2630
Card	-0.0029	1.4087	-0.0020	0.9984	-2.9287	2.9149

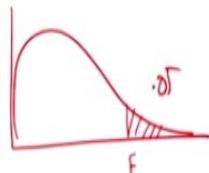
As I told you, you see that when you go for I have run this one previously, see here it is R square 0.783, here it is pseudo R square and as I told you in the previous class, you see that here there is a LL null, when there is no independent variable corresponding the value of likelihood is this much, this is null model, this is the base model, so this is your model likelihood, so we have to find the difference of these 2 to get the R square.

(Refer Slide Time: 16:29)

Testing for overall significance

Linear regression

- F-test of model fit
- $F = \frac{MSR}{MSE}$



Logistic Regression

- G-test of model fit

$$G = -2 \ln \left[\frac{\text{likelihood without the variable}}{\text{likelihood with variable}} \right]$$

$$= -2[-5 - (-4)] \\ = -2[-1] = 2$$



Now, we will see how to test the overall significance of linear regression and logistic regression, to test the overall significance of a linear regression we know that the F value is nothing but MSR divided by MSE, mean regression sum of square divided by mean error sum of square, then

what we will do; we will, the error of distribution will follow like this, we will find out what is the say, suppose this is 0.05, we will get corresponding F value, this is our calculated F value.

With the calculated F value is lying on that side will reject null hypothesis, if it is lying on acceptance say, you will accept it but for logistic regression, the formula is -2LL likelihood without the variable divided by likelihood with variable, this value can be find out, see - 2, when we say log value, it is division is nothing but subtraction, so likelihood without the variable is - 5 minus; this minus for the log of division, likelihood with variable is your - 4. When we simplify $-5 + 4$, so we will get - 1, so + 2.

This G also follow chi square distribution; this square is a right skewed distribution, this is your G value, G value is 2 for the degrees of freedom is number of independent variable in a logistical regression, that number of independent variable is nothing but the degrees of freedom for the G value, this value you can get it from this is output of Python of logistic regression and the linear regression, this was the comparison.

(Refer Slide Time: 18:40)

Testing for significance

<p>Linear regression</p> <ul style="list-style-type: none"> t-test $H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$ <div style="border: 1px solid orange; padding: 10px; width: 100%;"> $t = \frac{\hat{\beta}_1 - \beta_1}{S_{\beta_1}}$ <p>where $S_{\beta_1} = \sqrt{\frac{SS_{xx}}{n-2}}$</p> $S_{\beta_1} = \sqrt{\frac{SSE}{n-2}}$ $SS_{xx} = \sum X^2 - \frac{(\sum X)^2}{n}$ $\hat{\beta}_1 = \text{the hypothesized slope}$ $df = n - 2$ </div>	<p>Logistic regression</p> <ul style="list-style-type: none"> Wald-test <div style="border: 1px solid orange; padding: 10px; width: 100%;"> $W = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} = \frac{-0.0029}{1.4882}$ </div>
--	---

The another point here is to test the significance of each independent variable in a linear regression model, will use t test, the t test; the calculated t test is $b_1 - \beta_1$, here the assumption was beta 1 equal to 0 that was our null hypothesis, H_1 is beta 1 not equal to 0, then we will find

out t value, then we look for see $n - 2$ degrees of freedom, $n - k - 2$ degrees of freedom, then we will compare it, whether it lying on the acceptance in the rejection site.

The equivalent test in the logistic regression to test the individual significance of each independent variable we should go for this test called Wald test, this Wald test is nothing but estimated beta 1 divided by standard error of a beta 1, so here you go back, see here the for example, the card; the estimated beta 1 is -0.0029 , the standard error is this value is standard error 1.4887 , so this is equivalent to your z value.

So, this z value will be used to test whether the model is the individual independent variable is significant or not, this z value. This z value you got it, this dividing -0.0027 by 1.480 , this value is nothing but your Wald statistic.

(Refer Slide Time: 20:36)

Model Estimation fit

- The basic measure of how well the maximum likelihood estimation procedure fits is the likelihood value, similar to the sums of squares values used in multiple regression SSE
- Logistic regression measures model estimation fit with the value of -2 times the log of the likelihood value, referred to as $-2LL$ or $-2 \log \text{likelihood}$
- The minimum value for $-2LL$ is 0 , which corresponds to a perfect fit ($\text{likelihood} = 1$ and $-2LL$ is then 0)

Model estimation fit; the basic measure of how well the maximum likelihood estimation procedure fits is the likelihood value, similar to the sum of square value used in the multiple regressions, it is equivalent to your SSE. What will happen in multiple regression if the value of SSE is low, it is a good model, the same way logistic regression measures model estimation fit with the value of -2 times log of likelihood value referred to as $-2LL$.

The minimum value of $-2LL$ is 0, the similar to the SSE equal to 0, which corresponds to a perfect fit, so in the linear regression it is SSE, in the logistic regression it is $-2LL$, always we prefer lower is better.

(Refer Slide Time: 21:28)

Model Estimation fit

- The lower the $-2LL$ value, the better the fit of the model
- The $-2LL$ value can be used to compare equations for the change in fit

The lower the $-2LL$ value, the better fit the model is, the $-2LL$ value can be used to compare equations for change in fit. So, what will happen; first we have to run this model without any independent variable, we have to get what is $-2LL$, then we have to introduce another independent variable, then we have to compare how much error term is there. If it is lesser, then the variable which have included is explaining the model in better way that is a meaning of comparing equations for change in fit.

(Refer Slide Time: 22:07)

Between Model Comparison

- The likelihood value can be compared between equations to assess the difference in predictive fit from one equation to another, with statistical tests for the significance of these differences
- The basic approach follows three steps:

As I told you between model comparison, the likelihood value can be compared between equations to assess the difference in predictive fit from one equation to another with a statistical test for significance of these differences. There are 3 step is there that to assess whether the model is fit or not after introducing a new variable.

(Refer Slide Time: 22:32)

Step 1 : Estimate a null model

- The first step is to calculate a null model, which acts as the baseline for making comparisons of improvement in model fit.
- The most common null model is one without any independent variables, which is similar to calculating the total sum of squares using only the mean in linear regression. \bar{y}
- The logic behind this form of null model is that it can act as a baseline against which any model containing independent variables can be compared.

The first step is; we have to estimate the null model, what is null model? The first step is to calculate a null model, which act as a baseline for making comparison of improvement in the model fit. The most common null model is one without any independent variables which is similar to calculating the total sum of square using only the mean linear regression, it is like you

know y bar. The logic behind this form of null model is that it can act as a baseline against which any model containing independent variable can be compared.

(Refer Slide Time: 23:16)

Step 2: Estimate the proposed model

- This model contains the independent variables to be included in the logistic regression model.
- This model fit will improve from the null model and result in a lower $-2LL$ value.
- Any number of proposed models can be estimated

Step 2 is estimate the proposed model, this model contains the independent variables to be included in the logistic regression model, this model fit will improve from the null model and result in lower $-2LL$ value, if the after including a new independent variable, if the value of $-2LL$ low, then the model is good model, any number of proposed model can be estimated this way.

(Refer Slide Time: 23:47)

Step 3: Assess $-2LL$ difference:

- The final step is to assess the statistical significance of the $-2LL$ value between the two models (null model versus proposed model).
- If the statistical tests support significant differences, then we can state that the set of independent variable(s) in the proposed model is significant in improving model estimation fit.

The third step is assessing the $-2LL$ difference, the final step is to assess the statistical significance of the $-2LL$ value between 2 models that is null model versus proposed model, if the statistical tests support significant differences, then we can state that the set of independent variables in the proposed model is significant in improving the model estimation fit.

(Refer Slide Time: 24:20)

Between model comparison

Linear regression

- SSE
- $= \sum(y_i - \hat{y}_i)^2$

Logistic Regression

- $-2LL$ of proposed model

Another difference between logistic and linear regression is SSE, this also I have explained in my previous slide. In linear regression we say SSE, in logistic regression we say $-2LL$ of a proposed model, if it is lesser then model is good.

(Refer Slide Time: 24:40)

Between model comparison

Linear Regression

- $SSR = \sum(y_i - \bar{y}_i)^2$
- $SST - SSE$

Logistic regression

- Difference between log likelihood
- $= 2LL_{null} - (2LL_{model})$

In the linear regression we say SSR; in the logistic regression we say difference between log likelihood of null model and the model after introducing the 1 independent variable.

(Refer Slide Time: 24:54)

Normality of Residual (Error)

Linear regression

- Normally distributed
- Linear regression assumes that residuals are approximately equal for all predicted dependent variable values

Logistic regression

- Binomially distributed
- Logistic regression does not need residuals to be equal for each level of the predicted dependent variable values

Another important assumption between linear and logistic regression is we can say difference with respect to error is; a linear regression model the error term follow normal distribution but in a logistic regression, the error term follow binomial distribution. Linear regression assumes that the residuals are approximately equal for all predicted dependent variable values. Logistic regression does not need residuals to be equal for each level of predicted dependent variables.

(Refer Slide Time: 25:26)

Estimation Methods

- | | |
|--|--|
| <ul style="list-style-type: none">• Linear regression is based on <u>least square estimation</u>• Regression coefficients should be chosen in such a way that it minimizes the sum of the squared distances of each observed response to its fitted value | <ul style="list-style-type: none">• logistic regression is based on <u>Maximum Likelihood Estimation</u>• Coefficients should be chosen in such a way that it maximizes the Probability of Y given X (likelihood)• With MLE, the computer uses different "iterations" in which it tries different solutions until it gets the maximum likelihood estimates |
|--|--|

S

Another important difference is linear regression is based on the least square estimation via less but the logistic regression is based on the maximum likelihood estimation, this should be our first point. Regression coefficient should be chosen in such a way that it minimises the sum of the square distance of each observed responses to its fitted value, nothing but the error sum of square has to be minimised.

But here, the coefficient should be chosen in such a way that it maximises the probability of y given x , with the maximum likelihood estimation, the computer uses different iterations in which it tries to different solutions until it gets the maximum likelihood estimations, that is how many time solving logistic regression with the help of hand is very difficult.

(Refer Slide Time: 26:25)

Interpretation	
Coefficients of linear regression is interpreted as: <ul style="list-style-type: none">• Keeping all other independent variables constant, how much the dependent variable is expected to increase/decrease with an unit increase in the independent variable	In logistic regression, we interpret odd ratios as: <ul style="list-style-type: none">• The effect of a one unit of change in X in the predicted odds ratio with the other variables in the model held constant
	3

Another difference between logistic and linear regression is the way we interpret the coefficient; this is a very important difference. In logistic regression keeping all other independent variable constant, how much the dependent variable is expected to increase or decrease with an unit increase in the independent variable, this is the way we interpret the meaning of coefficient of each independent variable in a linear regression model.

But in a logistic regression model, the effect of 1 unit change in the X in the predicted odd ratio with respect to other variables in the model held constant, the point here is that the coefficient in the logistic regression is explained with the help of odd ratio, suppose the odd ratio is 3, if the

odd ratio is 3, if there is a 1 unit increase in the independent variable, there is a 3 times more chance, the probability will be increased.

This is the way to interpret the coefficient of logistic regression, in this class I have explained the difference between logistic regression and linear regression model, there are many parameters I have compared the equivalent term for logistic regression, equivalent means with respect to linear regression. If you are so thorough on interpreting the linear regression model after listening to this lecture, you can interpret the logistic regression model in a very easy manner. Thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology - Roorkee

Lecture – 41
Confusion Matrix and ROC

In this class, we are going to talk about how to check the performance of a logistic regression model. There are 2 ways to do that one; one is checking confusion matrix, another one is ROC, we will explain what is this confusion matrix and ROC, then we are going to see how we can check using these 2 criteria that the model which we developed is good or not.

(Refer Slide Time: 00:54)

Agenda

- Confusion matrix
- Receiver operating characteristics curve

The agenda for this class is we will see what is is confusion matrix and receiver operating characteristics curve.

(Refer Slide Time: 01:01)

Why Evaluate?

- Multiple methods are available to classify or predict
- For each method, multiple choices are available for settings
- To choose best model, need to assess each model's performance

Because we have seen in our previous example, there may be a different method to classify a set of data set. One of the methods is our logistic regression that is used to classify into 2 category, whether it is 0 category or 1 category but we want to see which method is the best one, so multiple methods are available to classify or predict. For each method, multiple choices are available for setting.

Here, multiple choices means that threshold value which we are going to say that beyond this probability, you should go to 1, below this probability you should come to 0, so that is our multiple choices. So, we have to know to choose the best model, we have to assess each model's performance that we will see in this class.

(Refer Slide Time: 01:53)

Accuracy Measures (Classification)

Misclassification error

(0, 1)

- Error = classifying a record as belonging to one class when it belongs to another class.
- Error rate = percent of misclassified records out of the total records in the validation data

In the classification context, how to measure the accuracy; one term is misclassification error, first we will see what is error. Error is classifying a record as labelling into one class when it belongs to another class. Suppose, when we say 0 to 1; 0, 1, there are 2 category; we can predicted, sometime what will happen, we may wrongly predicted, instead of saying 1, we will say it as 0, instead of saying 0, we will say it 1, so that is error. Error rate is percentage of misclassified records out of the total records in the validation data that is an error rate.

(Refer Slide Time: 02:42)

Confusion Matrix

Classification Confusion Matrix		
		Predicted Class
Actual Class	1	0
1	201	85
0	25	2689

201 1's correctly classified as "1"

85 1's incorrectly classified as "0"

25 0's incorrectly classified as "1"

2689 0's correctly classified as "0"

This is an example of confusion matrix; you see in row there is a actual class, in column, there is a predicted class. So, in row we see 1 0, 1 0, if you can predict 1 1 that is a correct one, actual also 1, the predicted value also 1, so like that we got this many number of data set. The other

possibility; the actual is 0, the predicted also 0, so these 2 columns, 2 cells are the correct value. So, here the frequency of correct saying 1, when it is actually 1 is 201.

The frequency of saying 0, when actually 0 is 2689, so the 201; 1 is correctly classified as 1, here the 85, 1 is incorrectly classified as 0, actual is 1 but we are predicting 0 that is your 85, the 25 represents incorrectly classified as 1, actually it is 0 but we classified as 1, 2689's are classified as 0, actual also 0, the predicted value is 0, so this is the set up for confusion matrix. This matrix is useful to find out the accuracy of our regression model.

(Refer Slide Time: 04:07)

Error Rate		
		Predicted Class
Actual Class	1	0
1	201	85
0	25	2689

Overall error rate = $(25+85)/3000 = 3.67\%$
Accuracy = $1 - \text{err} = (201+2689) = \underline{\underline{96.33\%}}$

If multiple classes, error rate is:
 $(\text{sum of misclassified records})/(\text{total records})$

We go here, how to find out the error rate; from error rate, we will see how to find the accuracy of our predicted model. See the overall error rate is; there are 2 error possibility, this 25 and 85, when you add this 25 + 85, the overall data set; overall count is 3000, so the error rate is possible error divided by 3000, so 3.67, the accuracy is 1 – error rate, so 1 minus that will give the 96.33%. If multiple classes is there, here only there are 2 classes there, one is 1 0. Sometimes there will be possibility of 2 also; in that case the error rate is sum of misclassified records divided by total records.

(Refer Slide Time: 05:02)

Cutoff for classification

Most algorithms classify via a 2-step process:

For each record,

1. Compute **probability of belonging to class "1"**
2. Compare to cutoff value, and classify accordingly

- Default cutoff value is 0.50
 - If ≥ 0.50 , classify as "1"
 - If < 0.50 , classify as "0"
- Can use different cutoff values
- Typically, error rate is lowest for cutoff = 0.50

Here, we will see cut-off for classification, so we need to have cut off to say when it is 1, when it is 0, most algorithms classifying via 2 step process. For each record, compute the probability of belonging to class 1, compare the cut off value and classify accordingly. The default cut-off value is 0.5, if the cut-off value is greater than or equal to 0.5, we will classify as 1. If the cut-off value is less than 0.5, we can classify as 0, okay.

In the probability range, we can have below 0.5, we say it is 0, above 0.5 is 1, this is the default value, we can use different cut off values, typically error rate is lowest for cut off, when you take the cut off value is 0.5.

(Refer Slide Time: 05:58)

Cutoff Table

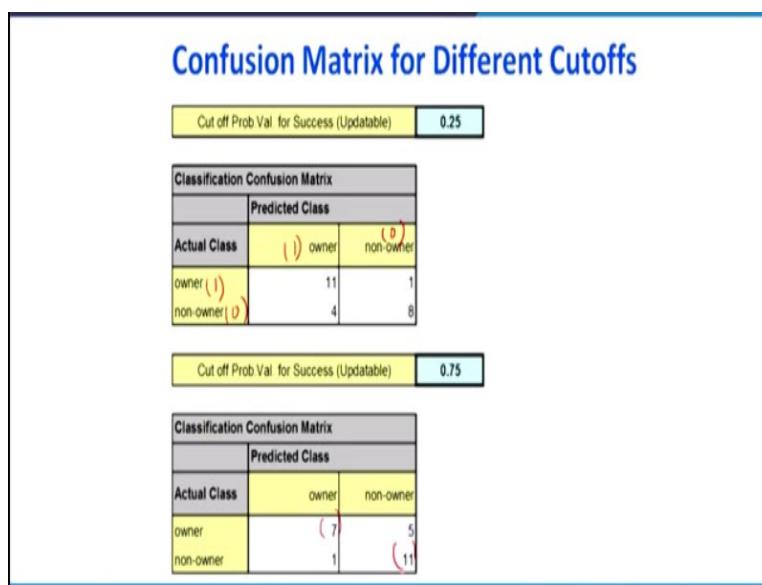
Actual Class	Prob. of "1"	Actual Class	Prob. of "1"
1	0.996	1	0.506
1	0.988	0	0.471
1	0.984	0	0.337
1	0.980	1	0.218
1	0.948	0	0.199
1	0.889	0	0.149
1	0.848	0	0.048
0	0.762	0	0.038
1	0.707	0	0.025
1	0.681	0	0.022
1	0.656	0	0.016
0	0.622	0	0.004

- If cutoff is 0.50: 11 records are classified as "1"
- If cutoff is 0.80: seven records are classified as "1"

For example, look at this picture, this is one example of our say, logistic regression model, this is our estimated y value. As I told you in our previous classes, the estimated y value is a probability, 0.996, 0.988 up to this is the continuing of this one. Suppose, if you keep cut-off is 0.5, what is the cut-off 0.5; 0.5 and above we are going to call it as 1, so this category. When you keep the cut-off here, there are 11 records classified as 1; 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11.

When you cut-off is 0.8, suppose when you put a cut-off here, 7 records are classified as 1, where 1, 2, 3, 4, 5, 6, 7, the problem comes what should be the right cut-off to classify it is 1 or 0 that is in our hand. Sometime, if you keep very high cut-off that also not good, if you keep very low cut-off that also not good that we will see, what is the meaning of keeping higher cut-off, what is the meaning of having lower cut-off?

(Refer Slide Time: 07:19)



Assume that my cut-off value is 0.25 in our previous problem, so cut-off this can be updated, mini software packages can be used we can keep different cut-off, so when you keep cut-off is 0.25, say this value, the actual is owner, the predicted may be 1 also, you can call it as 1, this is 1, this is 0, this is 0. So, 1 1 is the correct prediction, 0 0 it is correct prediction when you are keeping it is a 0.25. Suppose, you increase the cut-off value to 0.75, what will happen; we are able to predict only 7, here able to predict 11. So, what is happening; when you update the cut off, we are getting different confusion matrix that confusion matrix, every confusion matrix we will say about the overall accuracy.

(Refer Slide Time: 08:24)

Compute Outcome Measures

Confusion Matrix:

	Predicted Class = 0	Predicted Class = 1
Actual Class = 0	True Negatives (TN)	False Positives (FP)
Actual Class = 1	False Negatives (FN)	True Positives (TP)

N = number of observations

$$\text{Overall accuracy} = (\text{TN} + \text{TP})/\text{N}$$

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \quad (1) \quad \text{False Negative Error Rate} = \text{FN}/(\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \quad (2) \quad \text{False Positive Error Rate} = \text{FP}/(\text{TN} + \text{FP})$$

From the confusion matrix, generally the custom is first write 0 then 1, here also 0 1, you see here actual also 0, the predicted value also 0, it is a true negative, you see this diagonal value, actual also 1, predicted value also 1, so it is true positive. Whenever we do this mistake, what is happening here the actual is 0 but we are shown it is 1, so it is a false positive. I have some example in coming slides, what is the meaning of false positive, intuitively you can understand.

Similarly, the actual is 1 but you are predicting 0 that is your false negative, so these 4 cells are used to find out there are different parameter to check the prediction power of our regression model. The overall accuracy; these 2 cells are correct values TN, true negative plus true positive divided by sum of all cells value that is overall accuracy. The second point is sensitivity; sensitivity is true positive divide by true positive plus false negative that is sensitivity.

Because why we call it a sensitivity; actual also 1, predicted value also 1, so that is a sensitivity, so here the context of sensitivity; sensitivity of a testing machine that I will show you in the next slide. Then, specificity, specificity is true negative divided by true negative and false positive that is specificity. Here if you are predicting 0, we call it as specificity, if you are predicting 1 in a right way we are calling it a sensitivity.

Then the next term is overall error rate, what is an overall error rate; the false positive is one error plus false negative is another error divided by total number of elements. False negative error rate, where this is a false negative divided by true positive plus false negative that is a false negative error rate. False positive error rate, false positive divided by true negative plus false positive. I will explain what is the meaning of false positive, false negative in coming slides.

(Refer Slide Time: 10:59)

When One Class is More Important

In many cases it is more important to identify members of one class

- Tax fraud
- Credit default
- Response to promotional offer (0,1)
- Detecting electronic network intrusion
- Predicting delayed flights (1,0)

In such cases, we are willing to tolerate greater overall error, in return for better identifying the important class for further attention

Many times, the accuracy of the model is not important, sometime we may say that the one class is more important for example, predicting 1, when it is actually 1 that is more important. In many cases, it is more important to identify members of one class whether it is 0 or 1 but many time it is 1, 1 means when actual is predicted, actually it is 1, the predicted value also 1, so that is our more important class, we are not bother about when the 0 is predicted as 0, that is not important.

If it is 1, we should predicted as 1, so that time, the only one level is more important, for example tax fraud, credit default, response to promotional offer, detecting electrical network intrusion, predicting delayed flights, so there is a 2 possibility there, a person has done tax fraud or not, credit fault; default there are 2 possibility, this fellow will default or not. Response to promotional offer; whether this person will take the promotion offer or not.

If it is not taking no problem but we are considered about whether he is going to take the promotion offer or not because only between 0 and 1, we are more focus on 1, 0 is not important

for us, detecting electronic network intrusion, predicting delayed flight, whether the flight will be delayed or on time, so we are sometime we concerned about only the on time. In such cases, we are willing to tolerate greater overall error, in return for better identifying the important class for further attention. So, when you want to focus only one class out of these 2, that time accuracy is not important, something else important that will say.

(Refer Slide Time: 12:52)

ROC curves

- *ROC = Receiver Operating Characteristic*
- Started in electronic signal detection theory (1940s - 1950s)
- Has become very popular in biomedical applications, particularly radiology and imaging
- Also used in machine learning applications to assess classifiers
- Can be used to compare tests/procedures

That is done with the help of this curve called ROC curve, ROC is receiver operating characteristic curve, this curve is used to identify what should be the our threshold value to decide whether this category belongs to 1, whether this category belongs to 0, it was the idea started in electronic signal detection theory in 1940s to 1950s. It has become very popular in biomedical applications particularly, radiology and imaging.

Because if you want to predict a person having a disease or not, so this ROC is more suitable to decide whether there is a difference of different test, also used in machine learning application to assess classifier, in this class this ROC curve is used to decide or to evaluate whether the classifier is correctly classifying or not. Even it can be used to compare test or procedures here in the context of medical. So, what kind of operation can be done so, what kind of operation is more suitable for the patients?

(Refer Slide Time: 14:10)

ROC curves: simplest case

- Consider diagnostic test for a disease
- Test has 2 possible outcomes:
 - 'positive' = suggesting presence of disease
 - 'negative'
- An individual can test either positive or negative for the disease

We will see one example simple case; consider diagnostic test for a disease, you are asked to go for test, say medical test. The test has 2 possible outcomes; one is you may get positive that suggest that presence of disease, you may get negative that says absence of the disease. An individual can test either positive or negative for the disease. There are 2 possibility is there, a person may have the disease but you may get the negative report. Sometime a person may not have the disease but you may get positive report, so that is why the error started to come.

(Refer Slide Time: 14:59)

ROC Analysis

- **True Positives** = Test states you have the disease when you do have the disease
- **True Negatives** = Test states you do not have the disease when you do not have the disease
- **False Positives** = Test states you have the disease when you do not have the disease
- **False Negatives** = Test states you do not have the disease when you do

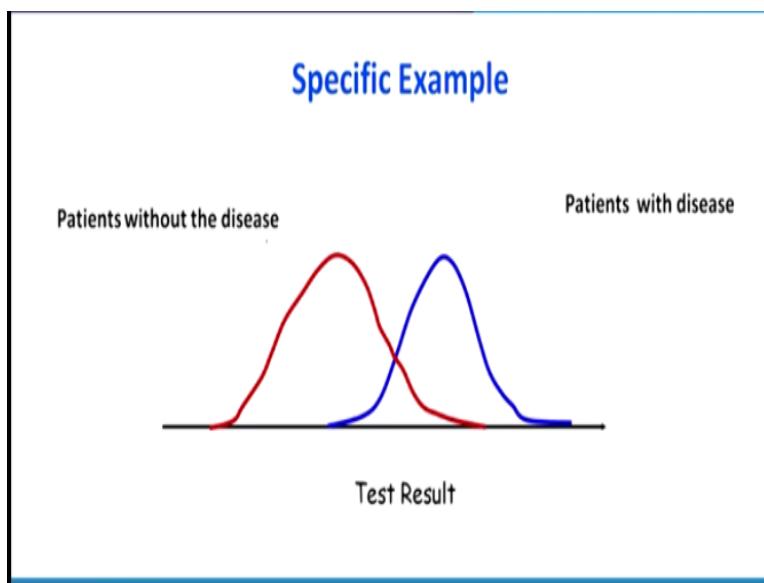
These terms you are going to use so often in coming slides, one is what is a true positive? Pictorially, I will show you in the next slide, the test states that you have the disease when you do have the disease, that means the person having the disease correctly it is saying that yes, you

have the disease. The true negative means the test states that you do not have the disease, when you do not have the disease, this is also no problem, you do not have disease, the report also; the test also says that you do not have disease.

The problem comes here in the false positive; the test states that you have the disease but when you do not have the disease, what is the meaning is that actually you do not have the disease but you shows the positive, positive means that you are saying that there is a disease, this is very dangerous that means, you do not have disease but the test that machine says is no, you have the disease.

Then the doctor started to, start the medication that may be dangerous also, there may be another category, test states that you do not have the disease, when you do, this also very dangerous actually, you have the disease but the test says you do not have the disease that this fellow may not get the proper medications because the test told that you do not have the disease, so both false positive and false negatives are dangerous.

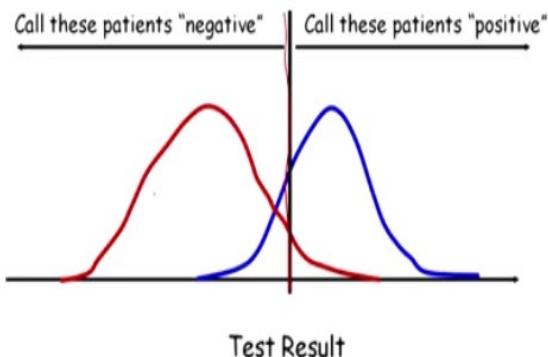
(Refer Slide Time: 16:34)



Look at this picture; the red colour shows that patients without the disease, the blue colour shows that patients with disease. Now, this is the test result, there are 2 possibilities there, a person without the disease, with disease.

(Refer Slide Time: 16:55)

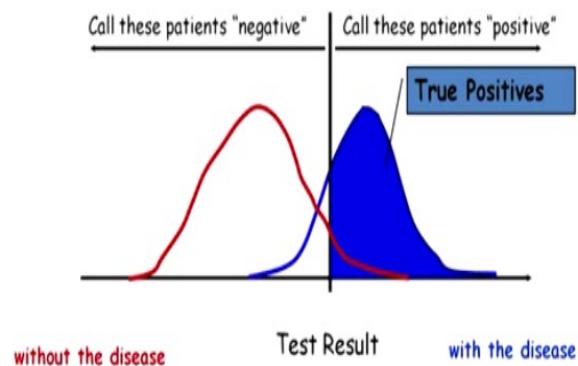
Threshold



Now, if you keep a cut-off here like this, you see that this in the x-axis shows kind of a probability. So, beyond this right hand side, you can call the patients having disease say, positive, beyond the left hand side of this line, you are going to say that the test shows negative that means that the patient is not having any disease.

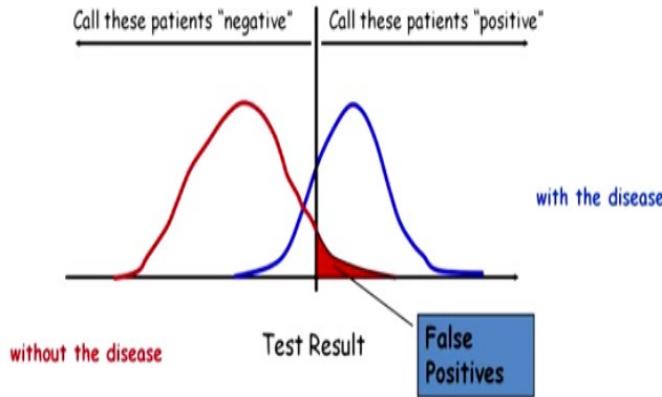
(Refer Slide Time: 17:21)

Some definitions ...



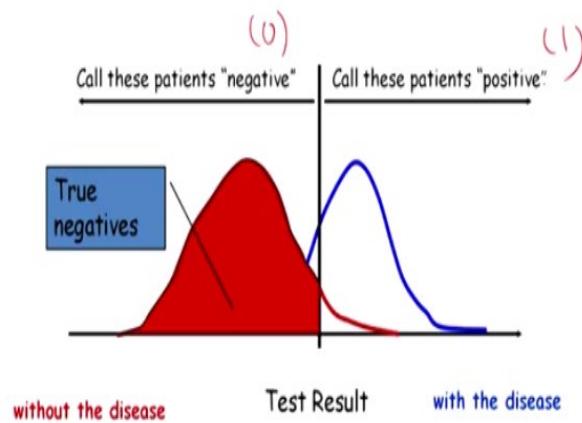
Now, you see that the blue one, this portion says true positive, what is a true positive? The person also actually have the disease, the test also says, yes you have the disease. Suppose, say for example 1 1, so that is a true positive.

(Refer Slide Time: 17:39)



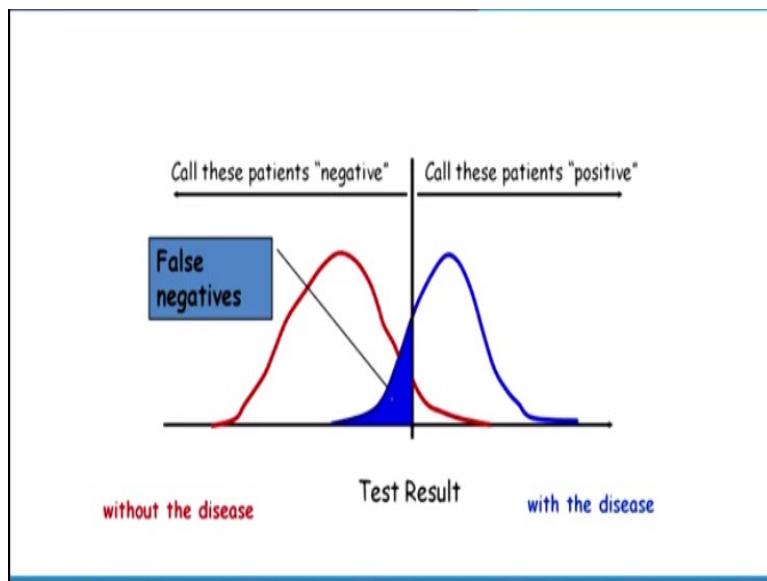
Now, look at this because the both negative and positive there is an overlapping, see the red portions actually, this red portions actually belongs to negative just because it is lying on positive side of this curve, we are going to say it is a false positive. False positive means actually, he is not having disease because of this cut-off which we have chosen it is lying on the positive side, we are going to say false positive, this is not good.

(Refer Slide Time: 18:26)



Because a person is not having disease but you are going to say is a disease, then you see another category true negative. When there is a cut-off, the left hand side portion says that true negative means the person not having disease, the test also says not having disease, it is like 0 0, 0, you code it to 0, no disease, this is disease, this also no problem because we will not bother about.

(Refer Slide Time: 18:50)

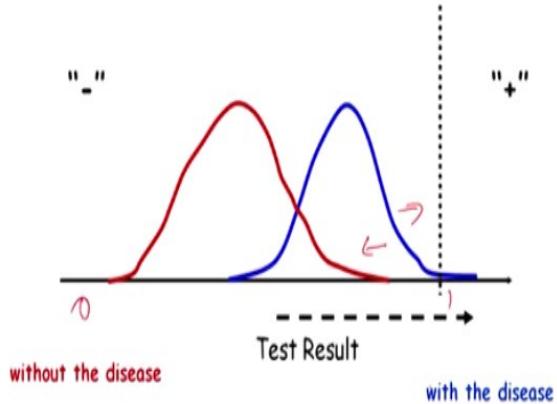


Now, what will happen in this case, this case actually is a false negative, actually this much portions of the blue they have the disease but it is lying on the negative side of the curve, we are going to say it is a false negative. The very common example for this one is sometime people may have confusion that person is having heart attack or the gastric trouble, so what will happen sometimes this is the false negative.

Actually, he had the heart attack, some people may suggest no, no, it is due to gas, so this is a false negative, this also very dangerous. Now, the question comes what should be the cut-off, suppose if you increase this cut-off what will happen?

(Refer Slide Time: 19:36)

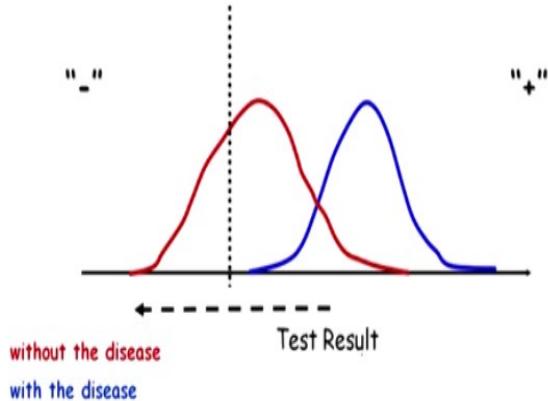
Moving the Threshold: right



That was the next one, suppose we have increase this cut-off like this, so what will happen when you increase the cut-off, so whoever comes there, we will give the report that negative, right because of our, so below this point is negative, below this point is positive. So, whoever goes to the pathological department, they will get a report that you do not have any disease because we have kept higher cut-off. Now, let us see what will happen now, see this one, this is the suppose, since it is the probability 0 to 1, see that whoever goes there, they will get a report, negative report.

(Refer Slide Time: 20:21)

Moving the Threshold: left



Then we will see another category, suppose if we decrease the cut-off, what will happen, when you decrease the cut-off, whoever goes to the pathology laboratory, they will get a positive

report; positive report means you may not have the disease but the report is going to say that no, you have the disease, you have to start the treatment. So, what is happening, the cut-off value plays very vital role to decide to minimise these errors. So, in this class what we are going to do, we are going to say that what is a role of this cut-off value and the accuracy of our predicted model or the classification model.

(Refer Slide Time: 21:02)

Threshold Value

- The outcome of a logistic regression model is a probability
- Often, we want to make a binary prediction
- We can do this using a *threshold value* t
- If $P(y = 1) \geq t$, predict positive
 - If $P(y = 1) < t$, predict negative
 - What value should we pick for t ?

So, we have to have the right threshold value; threshold value means that the vertical line, where it has to be chosen, whether it has to be chosen right hand side or left hand side because see the outcome of a logistic regression model is a probability often we want to make a binary prediction whether it is 0 or 1. We can do this using a threshold value t , call it as t , above this threshold value we are going to predict it is a positive, below the threshold value we are going to say it is predictive. Now, what is happening, what value should be pick for t , what should be the value of the cut-off value.

(Refer Slide Time: 21:41)

Threshold Value

- Often selected based on which errors are “better”
- If t is **large**, predict positive rarely (when $P(y=1)$ is large)
 - More errors where we say negative, but it is actually positive
 - Detects patients who are negative
- If t is **small**, predict negative rarely (when $P(y=1)$ is small)
 - More errors where we say positive, but it is actually negative
 - Detects all patients who are positive
- With no preference between the errors, select $t = 0.5$
 - Predicts the more likely outcome

This cut-off value is chosen based on which error is better, there are 2 error we have seen; false positive, false negative. If t is large, which I was shown you previously, predict positive rarely that means, if the t is high, we say report always that it is negative, so more errors where we say negative but it is actually positive, so what will happen here this curve, when you keep the higher threshold value, you see this fellow is a positive, this fellow is really having the disease.

But since because you have chosen higher cut-off value, we are going to give a report saying that negative that also dangerous that is a case of when we choose higher value of t value. Similarly, when you go for lower value of t value, the person may not have the disease but you are going to give a report saying that he has the disease, so both are dangerous. Now, you see the second category; if t is small, predict negative rarely when P of y equal to small.

More errors where we say positive but it is actually negative because we shifted the line to extreme left hand side, so that fellow is a positive but actually, it is negative, he is not having the disease, it detects all patients who are positive, whoever goes to that laboratory, they will get a report of that saying that you have the disease. So, with no preference between errors, you can select t equal to 5%.

Suppose, if you are not knowing, you are not able to say the cost of that error false positive or false negative, you can keep α equal to 0.5, it predicts the more likely outcome, it is a very conservative way.

(Refer Slide Time: 23:33)

Selecting a Threshold Value

- Compare actual outcomes to predicted outcomes using a *confusion matrix* (*classification matrix*)

	Predicted = 0	Predicted = 1
Actual = 0	True Negatives (TN)	False Positives (FP)
Actual = 1	False Negatives (FN)	True Positives (TP)

Now, I have brought this saying what is a true negative, true positive, selecting the threshold value, compare actual outcomes to predicted outcomes using confusion matrix, this also I have shown you. See that 0 0 it is a true negative when this is a false positive actually, this person is not having disease but you have given a report saying that he has a disease, this is a false negative; here false negative is this fellow actually having the disease.

(Refer Slide Time: 24:07)

True disease state vs. Test result		
Test Disease	not rejected/accepted	rejected
No disease ($D = 0$)	specificity	Type I error (False +) α
Disease ($D = 1$)	Type II error (False -) β	Power $1 - \beta$; sensitivity

But we have given a report saying that it is negative, this says the false positive is nothing but your alpha type I error. Here these type II false negative is nothing but beta, it is your type II error, this power of test we used to say in hypothesis testing $1 - \beta$, it is called sensitivity. If it is actual also 0, the predicted also 0, we say it is specificity.

(Refer Slide Time: 24:34)

		Predicted Class	
		C_0	C_1
Actual Class	C_0	$n_{0,0}$ = number of C_0 cases classified correctly	$n_{0,1}$ = number of C_0 cases classified incorrectly as C_1
	C_1	$n_{1,0}$ = number of C_1 cases classified incorrectly as C_0	$n_{1,1}$ = number of C_1 cases classified correctly

You see that in the term of; in the form of matrix say, C_0, C_0 , so number of C_0 cases classified correctly, you come to this diagonal, C_1 and C_1 , $n_{1,1}$ equal to number of C_1 cases classified correctly, the error comes here. What it says, actually it is 0 but we have predicted as 1. Similarly, this error has come, actually it is 1, we predicted as 0, so this is the confusion matrix.

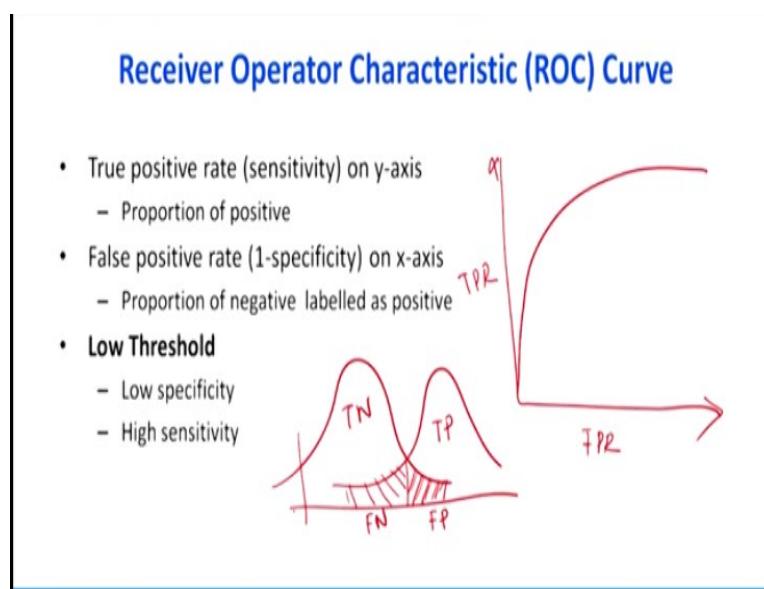
(Refer Slide Time: 25:08)

Alternate Accuracy Measures	
If " C_1 " is the important class,	
Sensitivity = % of " C_1 " class correctly classified	
Sensitivity = $n_{1,1} / (n_{1,0} + n_{1,1})$	
Specificity = % of " C_0 " class correctly classified	
Specificity = $n_{0,0} / (n_{0,0} + n_{0,1})$	
False positive rate = % of predicted " C_1 's" that were not " C_1 's"	
False negative rate = % of predicted " C_0 's" that were not " C_0 's"	

Now, let us explain this term what is sensitivity and specificity with respect to the previous table. If C1 is important class, the sensitivity equal to percentage of C1 class correctly classified, so sensitivity equal to $n_1\ 1$ divided by $n_1\ 0 + n_1\ 1$. Specificity equal to percentage of C0 class correctly classified, specificity equal to $n_0\ 0$ divided by $n_0\ 0 + n_0\ 1$, you can look at my previous slide; you can see that how it has come.

The, what is a false positive rate; percentage of predicted C1's that were not C1 that is a false positive rate, false negative rate is percentage of predicted C0's that were not C0 that is false positive, false negative rate because these terms we are going to use while constructing our ROC curve that is why I am defining what is a false; then we will say true positive and false positive, okay.

(Refer Slide Time: 26:23)



Then, what is a true positive rate; receiver operating characteristic curve, there will be, it will be go this way, in y axis, we have TPR, true positive rate. In x axis we will have 1 minus specificity that is false positive rate, so what will happen, receiver operating characteristic curve, the structure of ROC curve is in x axis, we will have true positive rate, this one, true positive rate that will be in y axis that is the proportion of positive cases.

In x axis, we are going to see false positive rate, what is a false positive rate; this false positive rate this portions which I shaded with this one, this is a false positive rate that we are going to

explain, this is 1 minus specificity, that will give your false positive rate, FPR. Now, what will happen; the curve will go this way, I will explain what will happen; when you keep very low threshold value, will have high sensitivity and low specificity.

Because low threshold in the sense, cut-off is here, there will be high sensitivity, so whoever comes to the laboratory, we will say that he has a disease, so the opposite of this is low specificity.

(Refer Slide Time: 28:44)

Selecting a Threshold using ROC

- Captures all thresholds simultaneously
- **High threshold**
 - High specificity
 - Low sensitivity
- **Low Threshold**
 - Low specificity
 - High sensitivity

The another category what will happen, when you keep higher threshold, it will have high specificity, very low sensitivity. So, whoever goes there will get a report saying that he is not having the disease, so high specificity. So, there is a contradiction of keeping higher cut-off value and lower cut-off value, so we have to choose the trade-off between the both the errors that we will see in the next class.

In this lecture, we have seen how to check the quality of our regression model, there are 2 methods; one is confusion matrix, another one is ROC analysis. I have explained using confusion matrix what is a difference cell means, what is an intuitive understanding of each cell that is what is a false positive and false negative, then I have explained some theory about the ROC analysis, then I have explained what will happen when the cut-off ratio is higher what will happen, when the cut-off ratio is very low, what will happen to that.

Now, in the next class we are going to see how to choose the correct cut-off value, so that in the next class in pictorially I will explain how to choose the correct cut-off value, so that there will be a trade-off between false positive and false negative error, thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology - Roorkee

Lecture – 42
Confusion Matrix and ROC - II

In this lecture, we will continue with our explanation of ROC analysis, in our previous lecture I have given you a theory about what is the confusion matrix and ROC analysis. In this lecture I will explain pictorially what is the different types of ROC curve, how that ROC curve is used to choose a correct classifying methodology that is ROC can help to predict the accuracy of our regression model.

(Refer Slide Time: 00:56)

Agenda

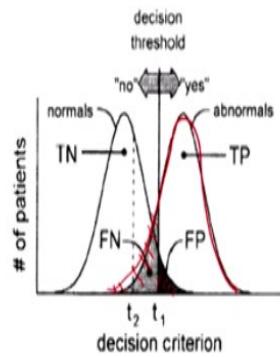
- Receiver operating characteristics curve
- Optimum threshold value

The agenda for this lecture is we will continue with receiver operating characteristic curve, then how to choose the optimal threshold value to classify the category whether it is 1 or 0, here we will use lot of pictures give you more understanding for you.

(Refer Slide Time: 01:12)

ROC analysis

- True Positive Fraction
 - $TPF = TP / (TP+FN)$
 - also called *sensitivity*
 - true abnormalities called abnormal by the observer
- False Positive Fraction
 - $FPF = FP / (FP+TN)$
- $Sensitivity = TN / (TN+FP)$
 - True normals called normal by the observer
 - $FPF = 1 - specificity$



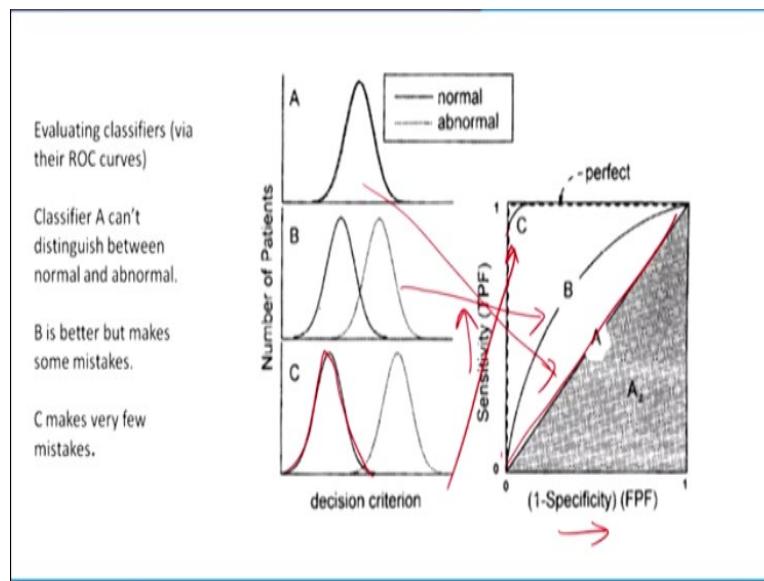
What is ROC analysis? As I told you the curve which I am showing previously, see this portion, this portion is abnormalities, person having disease, so if you are predicting exactly it is a true positive, sometime this much portions actually, he belongs to normals category, since it is lying on the abnormal side, we have given you report saying that you have the disease that is a false positive.

Now, this is the decision threshold, you see the another choice that whoever on the left hand side, we are going to say true negative, in the true negative sides, there are some people who belongs to positive side, their also lying on this side, in the negative side, so we are going to give a report false negative. What is a false negative? Even though they have the disease, we are going to say that you do not have the disease.

So, what is a true positive fraction; true positive fraction is true positive divided by your false negative, it is 1, also called sensitivity, true abnormalities called abnormal by the observer, this is the right way because if they are abnormal, we are going to say that yes, they are abnormal. False positive fraction is FPF, is a false positive divide by false positive plus true negative, then specificity; true negative divided by true negative and false positive.

True normals are called normals by the observer, this false positive rate is nothing but 1 minus specificity, so what will happen; false positive rate if you write 1 minus something you will end up with false positive divided by true negative plus false positive.

(Refer Slide Time: 03:23)



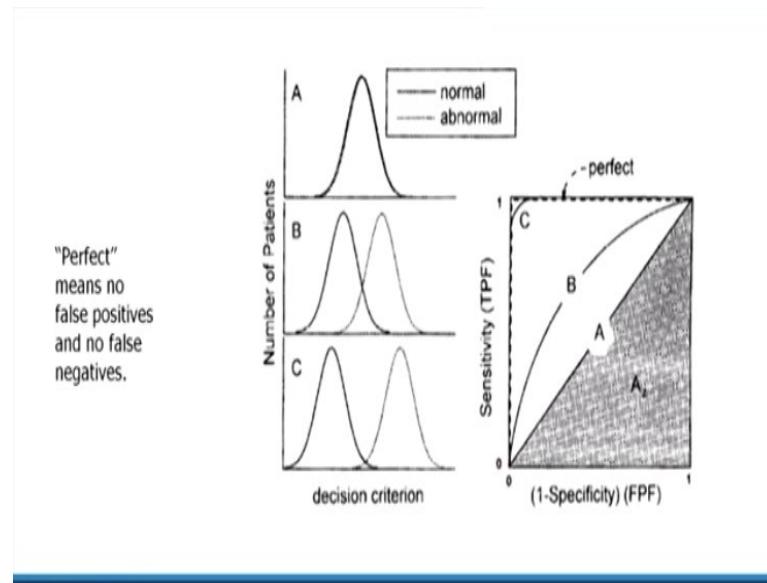
The next one is the true positive fraction, TPF equal to true positive, this curve is ROC curve as I told you in y axis, we have true positive fraction, in x axis we have false positive fraction, see the range is 0 0 because it is a problem it is 0 to 1, in y range also 0 to 1 1, this curve, the C curve you see that there is a C, you see that this category, when there is a somewhat overlap between this is a true negative, true positive, so this is the situation of your so this C.

So, we can I am writing here, so this was for this curve, you see that one there are some more, more overlap, so this is the situation when the B curves, you see there is a complete overlap, this is the situation of our A curve. If it is completely separated both true positive, true negative is 2 separate curves, so you will get a perfect ROC curve. Now, look at the different conditions, evaluating classifiers via ROC curves, classifier A, this one, this line cannot distinguish between normal and abnormal.

Because it is normal and abnormal there are 2 curves which are completely overlapped, the second one B is better but makes some mistakes, this situation because there is a somewhat overlap, C makes very few mistake, it is not completely separate, so this line, the reverse L shape

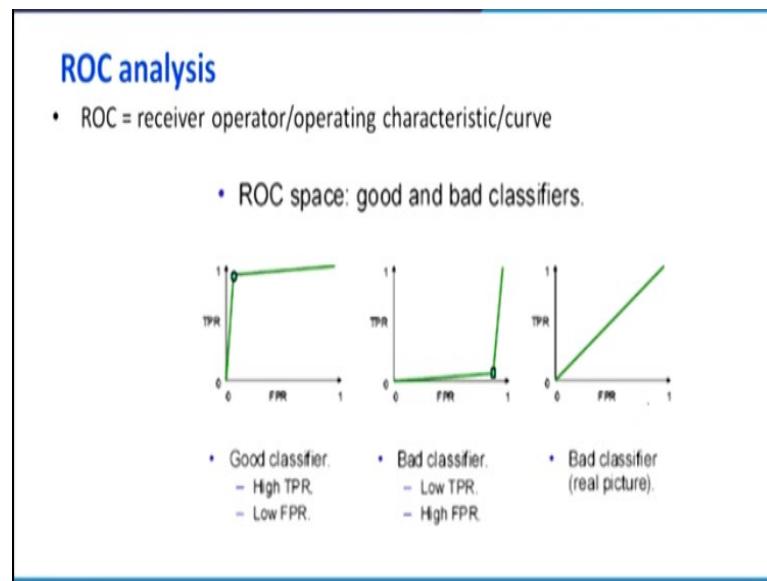
is perfect line. So, this graph, this picture connects between ROC curve and different types of your true positive and true negative.

(Refer Slide Time: 05:16)



There may be a perfect category, a perfect means no false positive and no false negative, so this line, no positive; no false positive and no false negative.

(Refer Slide Time: 05:32)

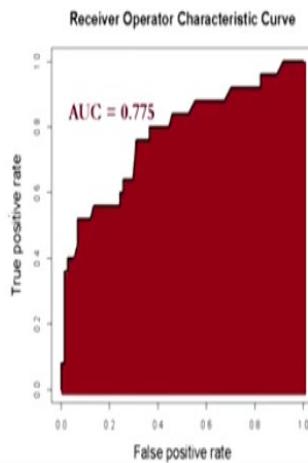


Look at there are different situations, see ROC curve; receiver operating or OC curve, ROC space good and bad classifier, you look at this one, this is a good classifier. Why we are saying it is a good classifier? High true positive rate and low false positive rate, this one we say bad

classifier because low true positive rate but high false positive rate, this one is a bad classifier because real picture because it is both are equal.

(Refer Slide Time: 06:10)

Area Under the ROC Curve (AUC)



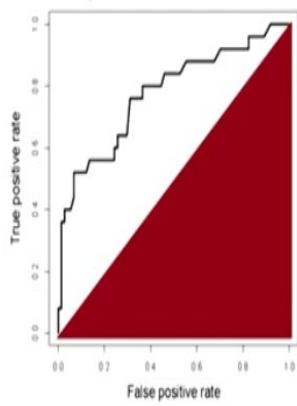
Next there is a one more term to explain the quality of a regression model that is area under curve, AUC that is nothing but in ROC curve, the area under ROC curve it is called AUC, area under curve. For example here AUC is 0.775, this portions where the red in colour, so all other things are same, the true positive rate, false positive rate.

(Refer Slide Time: 06:36)

Area Under the ROC Curve (AUC)

Receiver Operator Characteristic Curve

- What is a good AUC?
- Maximum of 1 (perfect prediction)
- Minimum of 0.5 (just guessing)



What is a good AUC, area under curve, see maximum it can hold everything that is a perfect prediction, so if the perfect prediction is there, that means, all 0's are predicted 0, all 1's are

predicted as 1, then you will get AUC, this full red colour. What is a good area under curve? The maximum value is 1 and minimum value is 0.5, it is just guessing one, so maximum it can go up to 1.

(Refer Slide Time: 07:09)

Selecting a Threshold using ROC

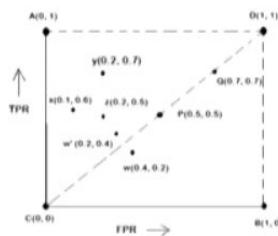
- Choose best threshold for best trade off
 - cost of failing to detect positives
 - costs of raising false alarms

Now, we will look for selecting a threshold using ROC curve, we have seen what is the different point of ROC curve, how to choose the threshold value that was important objective of this lecture. Choose the best threshold for best trade off, we are looking at cost of failing to detect positive and cost of raising false alarm, it is like a false positive and false negative, we have to see cost of that 2, whichever is more dangerous or more costly that should be minimised.

(Refer Slide Time: 07:42)

ROC Plot

- A typical look of ROC plot with few points in it is shown in the following figure.



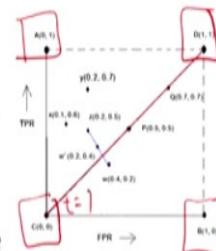
- Note the four cornered points are the four extreme cases of classifiers

Now, we will explain ROC plot each corners, a typical ROC plot with a few point in its shown in the following figure, there are 4 point is there; A, B, C, D, note that the 4 corner points are 4 extreme case of classifier, there are different points which are above the diagonal, some points are below the diagonal, we will take each and every points I will explain what is the significance of these points and how to interpret this point.

(Refer Slide Time: 08:17)

Interpretation of Different Points in ROC Plot

- The four points (A, B, C, and D)
- A: TPR = 1, FPR = 0, the ideal model, i.e., the perfect classifier, no false results
- B: TPR = 0, FPR = 1, the worst classifier, not able to predict a single instance
- C: TPR = 0, FPR = 0, the model predicts every instance to be a Negative class, i.e., it is an ultra-conservative classifier
- D: TPR = 1, FPR = 1, the model predicts every instance to be a Positive class, i.e., it is an ultra-liberal classifier



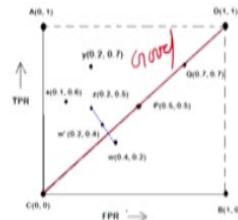
First, we will look at the point A; point A is this location, what is happening here; the true positive y value is 1, false positive is 0, this is an ideal model, the perfect classifier no false result, then we will go to the second category that is B, here the true positive rate is 0 because y value is 0 but x value is 1. The worst classifier not able to predict a single instance, then we will go for C, this situation where true positive also 0, false positive also 0.

The model predicts every instance to be negative class, it as an ultraconservative classifier, this will happen when t equal to 1. Suppose, if you keep a threshold that is very high level so everybody will be called it as negative class, the D this point is true positive rate also 1, false positive rate also 1, the model predict every instance to be positive class. When you take threshold value extreme left hand side so, whoever comes to the pathology laboratory, we will say that you have the disease, it is an ultra-liberal classifier.

(Refer Slide Time: 09:48)

Interpretation of Different Points in ROC Plot

- Let us interpret the different points in the ROC plot.
- The points on the upper diagonal region**
- All points, which reside on upper-diagonal region are corresponding to classifiers "good" as their TPR is as good as FPR (i.e., FPRs are lower than TPRs)
- Here, X is better than Z as X has higher TPR and lower FPR than Z.
- If we compare X and Y, neither classifier is superior to the other



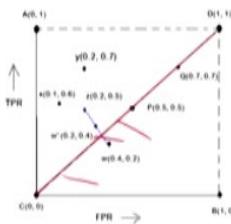
The problem comes how to choose the right ROC value, now look at different points inside the ROC curve. First we will look at the points on the upper diagonal region, all points which resides on upper diagonal region or corresponding to classifiers good because this portion is the good classifier, as their true positive rate is as good as false positive rate that is false positive rate is lower than the true positive rates.

See there is one point X, when you compare X and Z, X is better than Z because X has higher true positive rate and lower false positive rate than Z, when compare to X and Z, X is better. If you compare X and Y see that neither classifier is superior because there is a trade-off between TPR and FPR, if the TPR is increasing, FPR also increasing.

(Refer Slide Time: 10:55)

Interpretation of Different Points in ROC Plot

- Let us interpret the different points in the ROC plot.
- The points on the lower diagonal region
 - The Lower-diagonal triangle corresponds to the classifiers that are worst than random classifiers
 - A classifier that is worse than random guessing, simply by reversing its prediction, we can get good results.
 $W'(0.2, 0.4)$ is the better version than $W(0.4, 0.2)$, W' is a mirror reflection of W



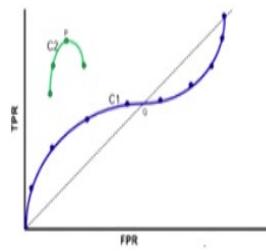
Now, let us interpret the different point ROC curve, first we will see the points on the lower diagonal region, previously I have explained the upper diagonal region, now we will look at the points in the lower diagonal region. The lower diagonal triangle corresponds to the classifier that are worse than the random classifier because this side it is not good because it is a high false positive rate.

A classifier that is worse than the random guessing simply by reversing its prediction, suppose look at the 2 point W dash and W , W dash is $0.2, 0.4$ is better version than the W 0.4 and 0.2 because W dash is the mirror reflection of W .

(Refer Slide Time: 11:41)

Tuning a Classifier through ROC Plot

- Using ROC plot, we can compare two or more classifiers by their TPR and FPR values and this plot also depicts the trade-off between TPR and FPR of a classifier.
- Examining ROC curves can give insights into the best way of tuning parameters of classifier.
- For example, in the curve C_2 , the result is degraded after the point P .
- Similarly for the observation C_1 , beyond Q the settings are not acceptable.



Now, tuning a classifier through ROC plot, see that I have 2 category of ROC plot, let us see which is better, why. Using ROC plot, we can compare 2 or more classifier by their TPR that is a true positive rate and false positive rate values and this plot also depicts the trade-off between true positive rate and false positive rate of a classifier. Examining ROC curves can give insight into the best way of tuning parameter of classifier.

For example, in this curve C2, the result is degraded after the point B, you see that this C2, what is happening; true positive rate is increasing after point B, what is happening; true positive rate is decreasing but there is no much decrease on false positive rate, so beyond this point P, it is not giving good classification. Similarly, for the observation C1, beyond Q, the setting are not acceptable because there is a comparatively lower false positive rate when compared to true positive rate.

(Refer Slide Time: 12:54)

Comparing Classifiers through ROC Plot

- We can use the concept of “area under curve” (AUC) as a better method to compare two or more classifiers.
- If a model is perfect, then its AUC = 1.
- If a model simply performs random guessing, then its AUC = 0.5
- A model that is strictly better than other, would have a larger value of AUC than the other.
- Here, C3 is best, and C2 is better than C1 as $AUC(C3) > AUC(C2) > AUC(C1)$.

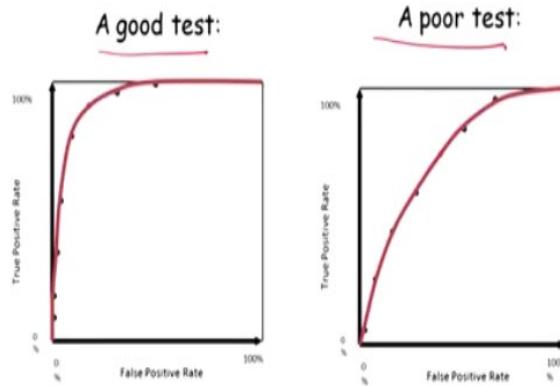
The figure shows an ROC plot with the x-axis labeled "FPR" and the y-axis labeled "TPR". Three curves are plotted: C1 (blue), C2 (red), and C3 (green). All curves start at the origin (0,0) and end at the top-right corner (1,1). Curve C3 is the uppermost, followed by C2, and then C1. A diagonal line from (0,0) to (1,1) represents a random classifier. The area under each curve represents the AUC, with C3 having the largest area and C1 the smallest.

Now, there are different classifying comparing different classifiers through ROC plot, when you look at this picture, see that there are C1, C2, C3 we can use the concept of area under curve as a better method to compare 2 or classifier, we can get different classifier by getting different threshold value. If a model is perfect, then the AUC is 1 which I have seen; which I have explained.

If a model simply performs a random guessing then the AUC is 0.5, so this area, a model that is strictly better than other would have larger value of AUC, area under curve than the other. So, out of these 3, the C3 is having higher area under curve, so that model and that corresponding threshold value is better to classify which is good or which is 1 or which is 0, when compared to C2, C1. Here, the C3 is the best, C2 is better than C1 as AUC, area under curve C3 is greater, then AUC , area under curve C2 greater than area under curve C1.

(Refer Slide Time: 14:17)

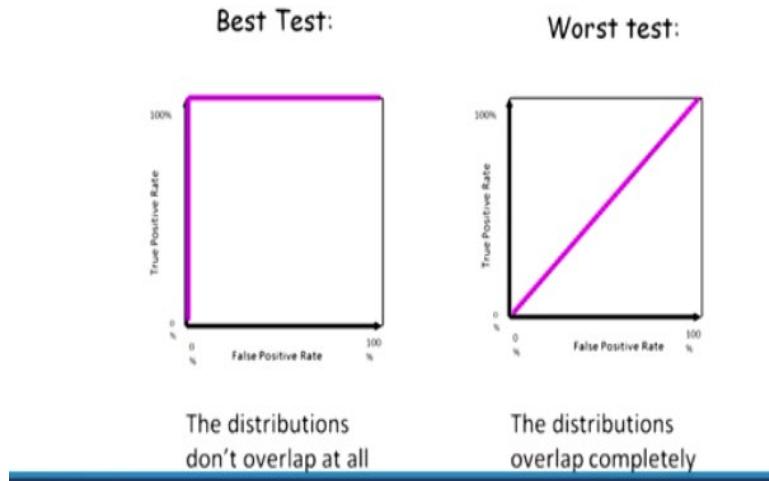
ROC curve comparison



Now, let us look at our extreme cases of this ROC curve, this was our typical ROC curve, see that how to compare 2 type of ROC curve, you see that it is closer, the area under curve is somewhat nearer to 1, so it is a good test. When you look at this one, area under curve of that ROC curve is lesser when compare to this, so it is a poor test. The left side one is the best test sorry, good test.

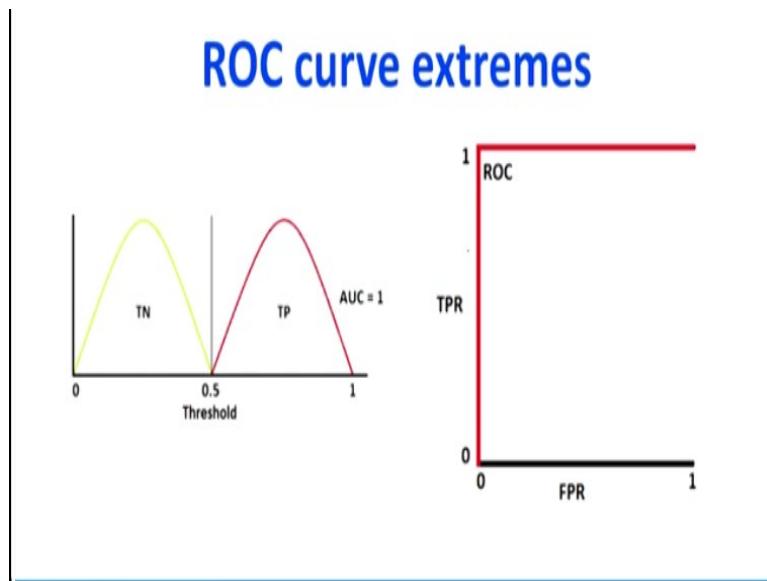
(Refer Slide Time: 14:55)

ROC curve extremes



Now, we will see extreme cases, when this extreme cases, see that the 2 distributions do not overlap at all, in the previous lecture I have shown you 2 cases, one is true negative and positive, this side true negative, true positive, then false positive, false negative. If the 2 lines are; 2 distributions are not overlapping, we will get very ideal case that is best test. The distributions will overlap completely, then you will get a this kind of diagonal, this is a worst test.

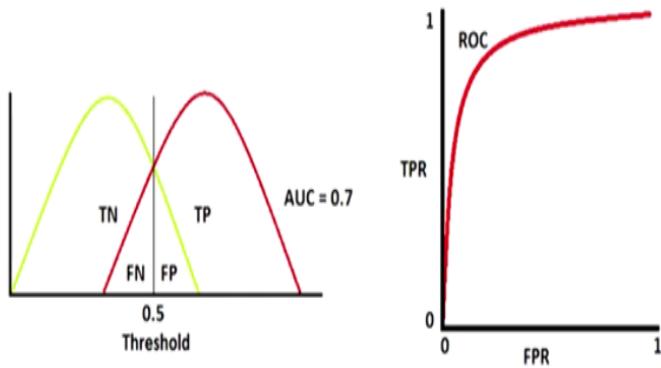
(Refer Slide Time: 15:32)



You see that this case, true negative true positive, there is no overlap at all, so you will get this kind of ROC curve.

(Refer Slide Time: 15:43)

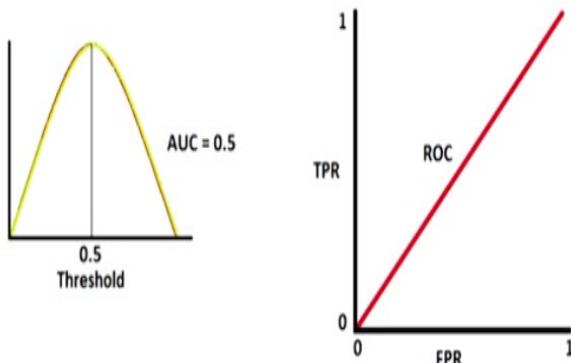
Typical ROC



You see that there are somewhat some overlap is there, so you will get a typical ROC curve because there is area under curve, previously it was 1, now it is area under curve is 0.7.

(Refer Slide Time: 15:57)

ROC curve extremes



Now, you see this case the area, both the distributions are completely overlapping that whenever it is overlapping, area under curve is 0.5, so corresponding ROC curve will look like this, so far we have seen and we have understood the concept of confusion matrix and ROC curve, now I am going to take one example, that example already I have discussed in my previous lectures. With the help of that example I am going to tell you how to choose the correct threshold value to classify whether it is belongs to category 1 or category 2.

(Refer Slide Time: 16:38)

Variables

- Management thinks that **annual spending** at Simmons Stores and whether a **customer has a Simmons credit card** are two variables that might be helpful in predicting whether a customer who receives the catalog will use the coupon.
- Simmons conducted a pilot study using a random sample of 50 Simmons credit card customers and 50 other customers who do not have a Simmons credit card.
- Simmons sent the catalog to each of the 100 customers selected.
- At the end of a test period, Simmons noted whether the customer used the coupon or not?

The example is this book; this example is taken from this book statistics for business and economics from David Anderson Sweeney and Williams. The example is; let us consider an application of logistic regression involving direct mail promotion being used by Simmons Stores. Simmons owns and operates a national chain of women's apparel stores, 5000 copies of an expensive 4 colour sales catalog have been printed and each catalog includes a coupon that provides 50 dollar discount on purchase of 200 dollar or more. The catalog are expensive and Simmons would like to send them to only those customers who have the highest probability of using the coupon.

(Refer Slide Time: 17:28)

Variables

- Management thinks that **annual spending** at Simmons Stores and whether a **customer has a Simmons credit card** are two variables that might be helpful in predicting whether a customer who receives the catalog will use the coupon.
- Simmons conducted a pilot study using a random sample of 50 Simmons credit card customers and 50 other customers who do not have a Simmons credit card.
- Simmons sent the catalog to each of the 100 customers selected.
- At the end of a test period, Simmons noted whether the customer used the coupon or not?

So, what are the variables which are involved in this problem is; one is annual spending, another one is whether the customer has Simmons credit card or not. What we are going to predict whether a customer who receives the catalog will use the coupon or not, Simmons conducted a pilot study using a random sample of 50 Simmons credit card customers and 50 customers who do not have Simmons credit card. Simmons sent the catalog to each of the 100 customer selected, at the end of the test period Simmons noted whether the customers used the coupon or not, this is the problem.

(Refer Slide Time: 18:11)

Data (10 customer out of 100)

Customer	Spending	Card	Coupon
1	2.291	1	0
2	3.215	1	0
3	2.135	1	0
4	3.924	0	0
5	2.528	1	0
6	2.473	0	1
7	2.384	0	0
8	7.076	0	0
9	1.182	1	1
10	3.345	0	0

For that problem this is the dataset, spending how much they spend in the last month that is our one of the independent variable. Possession of Simmons credit card, if it is 1 he has, 0 does not have that is the another independent variable. The coupon; whether he use the coupon or not that is our dependent variable.

(Refer Slide Time: 18:33)

Explanation of Variables

- The amount each customer spent last year at Simmons is shown in thousands of dollars and the credit card information has been coded as 1 if the customer has a Simmons credit card and 0 if not.
- In the Coupon column, a 1 is recorded if the sampled customer used the coupon and 0 if not.

You see that dependent variable is 2 category 0 or 1, so it is a; then we have to go for logistical regression. The amount of each customer spent last year at Simmons is shown in the 1000's of dollars and the credit card information has been coded as 1, if the customer has the Simmons credit card 0, if not. In the coupon column, 1 is recorded if the sampled customers used the coupon and 0 if not.

(Refer Slide Time: 19:04)

Loading data file and get some statistical detail

```
In [1]: 1 import pandas as pd
2 import matplotlib.pyplot as plt
In [2]: 1 data = pd.read_excel('Simmons.xls')
2 data.head()
Out[2]:
   Customer  Spending  Card  Coupon
0         1      2.291     1     0
1         2      3.215     1     0
2         3      2.135     1     0
3         4      3.924     0     0
4         5      2.528     1     0
In [3]: 1 data.describe() # it is used to get some statistical detail
Out[3]:
   Customer    Spending      Card      Coupon
count  100.000000  100.000000  100.000000  100.000000
mean   50.500000   3.333790   0.500000   0.402366
std    29.011492   1.741291   0.502519   0.492366
min    1.000000   1.058000   0.000000   0.000000
25%   25.750000   2.059000   0.000000   0.000000
50%   50.500000   2.865500   0.500000   0.000000
75%   75.250000   4.468250   1.000000   1.000000
max   100.000000   7.076000   1.000000   1.000000
```

So, we have imported the data, for import we have imported necessary libraries, import pandas as pd, import matplotlib.pyplot as plt, then the dataset is Simmons.xls, I am going to show, run this Python code and I am going to explain further. I brought the screenshot of my Python

output, so data dot head we came to know there is a spending is one independent variable, card and coupon.

The first one is a data dot describe, that is to get an idea about the details of each variables, customer, there is no meaning for this one, for spending you see that there are 100 values is there, the mean is 3.3, standard deviation is this one, since card is the categorical variable, there is no meaning, for the mean there is no meaning for standard deviation, so it is not applicable. The coupon also, it is a categorical variable, in the categorical variable there are 100 values is there. Here also, there is no meaning for mean and standard deviation because you cannot do any arithmetic operation when there is a nominal or categorical variable.

(Refer Slide Time: 20:17)

Method's description

- Dataframe.describe(): This method is used to get basic statistical details such as central tendency, dispersion and shape of dataset's distribution.
- Numpy.unique(): This method gives unique values in particular column.
- Series.value_counts(): Returns object containing counts of unique values.
- ravel(): It will return one dimensional array with all the input array elements.

We are going to use different inbuilt functions, so if you say, Dataframe dot describe that function is used to get the basic statistical details such as central tendency, dispersion and shape of the datasets distribution. If you use numpy.unique, this method gives the unique value in a particular column, they count this option; Series.value_counts return object containing count of unique values.

(Refer Slide Time: 21:01)

Split dataset into training and testing sets

```
In [4]: 1 data['Coupon'].unique() # It gives unique value in particular column
Out[4]: array([0, 1], dtype=int64)

In [5]: 1 data['Coupon'].value_counts()
Out[5]:
0    60
1    40
Name: Coupon, dtype: int64

In [7]: 1 from sklearn import linear_model
2 from sklearn.model_selection import train_test_split
3 from sklearn.linear_model import LogisticRegression

In [8]: 1 x = data[['Card','Spending']]
2 y = data['Coupon'].values.reshape(-1,1)
3 x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25, random_state = 42)

In [9]: 1 len(x_train), len(y_train), len(x_test), len(y_test)
Out[9]: (75, 75, 25, 25)
```

This ravel; it will return one dimensional array with all the input array elements, so I use that inbuilt function for example, data for the coupon column unique, so it is give 0, 1 because that means we can come to know there are 2 category in the coupon column, one is 0, another one is 1. Then how many when you say dot value underscore counts, say there are 60-0, 40-1, 60 people did not use the coupon, 40 people have use the coupon.

Then for running logistic regression and they split the data into 2 categories; some data for training and building the model, after the model is built, we will use the test data to verify our built model from sklearn import linear_model, from sklearn.model_selection import train_test_split, from sklearn.linear_model import LogisticRegression.

So, the x value equal to data; card and spending; independent variable, y value is coupon, okay, first we are going to split x underscore train, x underscore test that is x value after splitting we are going to call it this way then, y train y test equal to train underscore test underscore split x, y test underscore size 0.25, you can take any value, thus I have set some value, so that we can repeat the code again, you may get the same kind of output.

So, you see that I wanted to say see for the in the training data set, there is 75 data set for x and 75 data set for y. For testing data set, there are 25 data set for x, 25 data set for y, this is for testing purpose.

(Refer Slide Time: 22:50)

Building the model and predicting values

```
In [10]: 1 Lreg = LogisticRegression(solver='lbfgs')
2 Lreg.fit(x_train, y_train.ravel()) #ravel() will return 1D array with all the input-array elements

Out[10]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                           intercept_scaling=1, max_iter=100, multi_class='warn',
                           n_jobs=None, penalty='l2', random_state=None, solver='lbfgs',
                           tol=0.0001, verbose=0, warm_start=False)

In [11]: 1 y_predict = Lreg.predict(x_test)
2 y_predict

Out[11]: array([1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1,
               1, 0, 0], dtype=int64)

In [12]: 1 y_predict_train = Lreg.predict(x_train)
2 y_predict_train

Out[12]: array([0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1,
               0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
               0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
               0, 0, 0, 0, 0, 1, 1, 0], dtype=int64)
```

First, we will use logistic model and will after building the model, we will predict the values, we are going to use in the logistic regression, lbfgs, there are different method I will show you there, when I am running the code, this is one method for constructing the logistic regression model, logisticregression dot fit x train y train dot ravel, we will written one-dimensional array with all input array elements, so we are getting this output.

Now, we will predict the y value by using the test data set, what I have done; we have constructed the model, then we will use a test dataset to predict the y value, this was our predicted y value, then we will predict y value by using the training dataset because training we have 75 data set, for testing there are 25 dataset, so this was after predicted the regression model using training dataset.

(Refer Slide Time: 23:52)

Calculate probability of predicting data values

```
In [13]: 1 y_prob_train = Lreg.predict_proba(x_train)[:,1]
2 y_prob_train.reshape(1,-1)
Out[13]: array([0.49622117, 0.32880793, 0.44329114, 0.33320924, 0.41456465,
   0.32890329, 0.3975041, 0.66921229, 0.25844511, 0.63672372,
   0.29274306, 0.20466974, 0.5159296, 0.41992276, 0.24342356,
   0.528514, 0.47945107, 0.52805789, 0.33191449, 0.27457435,
   0.49179296, 0.63261616, 0.24690181, 0.47089452, 0.27842076,
   0.41663875, 0.36155602, 0.49970327, 0.23621636, 0.37860052,
   0.40809323, 0.20877877, 0.28563859, 0.37231802, 0.65309742,
   0.43807264, 0.33630478, 0.40406407, 0.23431177, 0.37202304,
   0.49970327, 0.39768396, 0.32880793, 0.25782412, 0.47393034,
   0.42870861, 0.26520939, 0.33320924, 0.54682499, 0.45446086,
   0.44326597, 0.4965167, 0.60065954, 0.3898954, 0.49149447,
   0.27414424, 0.27785686, 0.67464141, 0.28195004, 0.48593427,
   0.38633227, 0.31373449, 0.42810085, 0.27418723, 0.44371771,
   0.41629601, 0.642004, 0.6571001, 0.44068025, 0.28195004,
   0.40217015, 0.43807264, 0.50977653, 0.57944626, 0.2904233])]} 75 /
```

```
In [14]: 1 y_prob = Lreg.predict_proba(x_test)[:,1]
2 y_prob.reshape(1,-1)
3 y_prob
Out[14]: array([0.52802946, 0.49516653, 0.45703306, 0.27712052, 0.34390407,
   0.26025171, 0.272112052, 0.607686, 0.42836534, 0.43637155,
   0.31307455, 0.23676248, 0.45703306, 0.43602766, 0.37596116,
   0.44900317, 0.46952365, 0.68521935, 0.25167254, 0.47073304,
   0.42361093, 0.56580644, 0.52792177, 0.40302605, 0.27457351])]} 25
```

Next one; y underscore probability underscore train, here see that this is the probability value which dataset for training dataset, so there are 75 dataset, we have got the probability of all the 75 dataset. Our problem is going to be there, what should be our cut-off value here, to say this below this category is called 0, above that category is 1, so y underscore probability here also we will find out for using test dataset, we will predict the probability, this is there going to be 25 dataset, here going to be 75 dataset.

(Refer Slide Time: 24:34)

Summary for logistic model

```
In [15]: 1 x = data[['Spending', 'Card']]
2 y = data['Coupon']
3
4 import statsmodels.api as sm
5 xl = sm.add_constant(x)
6 logit_model=sm.Logit(y,xl)
7 result=logit_model.fit()
8 print(result.summary())
Optimization terminated successfully.
   Current function value: 0.604869
   Iterations 5
Logit Regression Results
Dep. Variable: Coupon No. Observations: 100
Model: Logit Df Residuals: 97
Method: MLE Df Model: 2
Date: Mon, 16 Sep 2019 Pseudo R-squ.: 0.1012
Time: 10:15:13 Log-Likelihood: -60.487
converged: True LL-Null: -67.301
LLR p-value: 0.001098
coef std err z P>|z| [0.025 0.975]
const -2.1464 0.577 -3.718 0.000 -3.278 -1.015
Spending 0.3416 0.129 2.655 0.008 0.089 0.594
Card 1.0987 0.445 2.471 0.013 0.227 1.970
```

This is that the; our task is what should be the cut-off value or threshold value, first we will construct a regression equation, logistic regression x is x data y data, after importing the model we got this one, constant is -2.1464, spending independent variable is 0.3416, card 1.089, when

we look at the pseudo R square, it is good, then the P value also good, the overall model is good, even if you look at the Wald test, so this P value also all are less than 0.05, this model is good.

(Refer Slide Time: 25:20)

Accuracy Checking

- By using accuracy_score function.
- By using confusion matrix

		Predicted (0)	Predicted (1)
Actual (0)	True Negative(tn)	False Positive(fp)	
	False Negative(fn)	True Positive(tp)	
Actual (1)			

The point is how to set the threshold value, so after getting this model, we have predicted some portions are 1, so first we will go for checking the accuracy. There are 4 possibility; the actual is 0, predicted is 0, it is a true negative, actual is 1, predicted is 1, it is a true positive, actual is 0, the predicted is 1, it is a false positive, actual is 1, predicted is 0 that is a false negative.

(Refer Slide Time: 25:44)

Calculating Accuracy Score using Confusion Matrix

```
In [16]: 1 from sklearn.metrics import accuracy_score  
2 score =accuracy_score(y_test,y_predict)  
3 score  
Out[16]: 0.76
```

```
In [17]: 1 from sklearn.metrics import confusion_matrix  
2 confusion_matrix(y_test, y_predict)  
Out[17]: array([[15,  1],  
   [ 5,  4]], dtype=int64)
```

```
In [18]: 1 tn, fp, fn, tp = confusion_matrix(y_test, y_predict).ravel()  
2 print("True Negatives: ",tn)  
3 print("False Positives: ",fp)  
4 print("False Negatives: ",fn)  
5 print("True Positives: ",tp)
```

True Negatives: 15
False Positives: 1
False Negatives: 5
True Positives: 4

Because that table is the; the previous table is the base for the confusion matrix, so here the accuracy of model is test by using this score function, so score equal to accuracy underscore

score for y test and y predicted, so the accuracy of that model is 76, to get the confusion matrix; confusion underscore matrix y underscore test y underscore predict, we are getting this confusion matrix. So, here what is meaning is the true negative is 15, true positive is 4, so the false positive is 1, false negative is 5, so this is the way to write the confusion matrix.

(Refer Slide Time: 26:31)

Generating Classification Report

```
In [19]: 1 from sklearn.metrics import classification_report
2 print(classification_report(y_test, y_predict))

precision    recall  f1-score   support
          0       0.75      0.94      0.83      16
          1       0.80      0.44      0.57       9
   avg / total       0.76      0.76      0.76      25
  macro avg       0.70      0.69      0.70      25
weighted avg     0.77      0.76      0.74      25
```

- Recall gives us an idea about when it's actually yes, how often does it predict yes.
- Precision tells us about when it predicts yes, how often is it correct

Generating classification report; you see that when you use this functional classification underscore report, we will get this output, this is there are different columns, see one is on precision, another one is a recall, another one is f1 score and support. This recall gives us an idea about when it is actually yes, how often it predicts yes, it is like our sensitivity. Precision tells us about when it is predict yes, how often is it correct.

I have explained what is the recall, so recall gives us an idea about when it is actually yes, how often does it predict yes, it take care both specificity and sensitivity. If it is 1, we call it sensitivity, if it is 0, we call it specificity. So, in our problem the specificity is 0.94 that says that see, 94% of time we got 0 and we have predicted also it is 0, when you say sensitivity 0.44, so 44% of time, the actual is 1, we predicted also 1.

(Refer Slide Time: 27:52)

Interpreting Classification Report

- Precision = $tp / (tp + fp)$
- Accuracy = $(tp + tn) / (tp + tn + fp + fn)$
- Recall = $tp / (tp + fn)$

	Predicted (0)	Predicted (1)
Actual (0)	tn	fp
Actual (1)	fn	tp

So, the next one more column is the precision; the precision tells us about when it predict yes, how often is it correct, will explained this one, meaning of precision in the next slide. Now, we will interpret the classification report which was the; in the previous slide I have shown that output. The precision is true positive divided by true positive plus false positive, the accuracy is here there is one; so this cell is accurate value, this cell also accurate value.

So, sum of these 2 divided by sum of all the cells, so recall is a true positive divide by true positive plus false negative. In this lecture I have explained graphically what is ROC curve and how to choose a ROC curve, with the help of some pictures. At the end, I have taken one problem, there in that problem I explained what is a confusion matrix, how to interpret each cell in the confusion matrix. In the next class, we will continue and I will explain how to choose ROC value with the help of this same example that we will see in the next class, thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 43
Performance of Logistic Model – III

In this lecture, we are going to test the performance of logistic regression model. We use Python and I will show you a demo how to check the performance of a logistic regression model.

(Refer Slide Time: 00:41)

Agenda

Python demo for accuracy prediction in logistic regression model using Receiver operating characteristics curve

The agenda for this lecture is Python demo for accuracy prediction in logistic regression model using ROC curve.

(Refer Slide Time: 01:11)

Sensitivity and Specificity

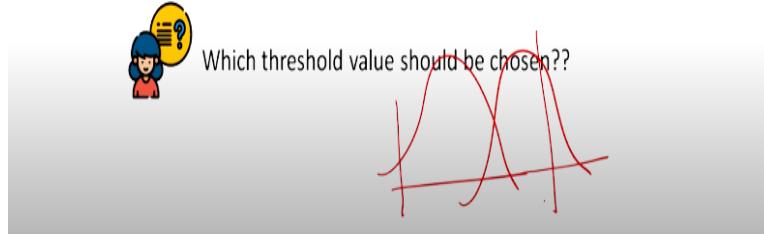
- For checking, what type of error we are making; we use two parameters-
1. Sensitivity = $tp/(tp+fn)$ → True Positive Rate(tpr)
 2. Specificity = $tn/(tn+fp)$ → True Negative Rate (tnr)

There are two terms, one is Sensitivity and another one is Specificity. For checking what type of error we are making, we use 2 parameter. One is sensitivity. The another name for sensitivity is True Positive Rate. This also I have shown you in your previous lecture tp divided true positive by false negative. Specificity is a true negative rate, that is true negative divided by true negative plus false positive.

(Refer Slide Time: 01:20)

Specificity and Sensitivity Relationship with Threshold

Threshold (Lower)	Sensitivity (\uparrow)	Specificity (\downarrow)
Threshold (Higher)	Sensitivity (\downarrow)	Specificity (\uparrow)



In this lecture, I am going to stay the connection between sensitivity and specificity for different threshold value. The first case is, when the threshold value is low, when the threshold value is suppose this way, suppose when you put threshold value here, we will increase the sensitivity, but decrease our specificity. When the threshold value is higher, suppose if you keep the threshold value here, what will happen specificity will increase, sensitivity will decrease. So, which threshold value should be chosen? That is the problem. That I will show you with the help of Python programming.

(Refer Slide Time: 02:04)

Measuring Accuracy, Specificity and Sensitivity

```
In [20]: 1 Accuracy = (tp + tn) / (tp + tn + fp + fn)
2 print("Accuracy {:.2f}".format(Accuracy))
Accuracy 0.76
```



```
In [21]: 1 Specificity = tn/(tn+fp)
2 print("Specificity {:.2f}".format(Specificity))
Specificity 0.94
```



```
In [22]: 1 Sensitivity = tp/(tp+fn)
2 print("Sensitivity {:.2f}".format(Sensitivity))
Sensitivity 0.44
```

$t=0.5$

First, we will check what accuracy is. Accuracy is true positive plus true negative divided by true positive plus true negative plus false positive plus false negative. So, the accuracy for our problem is 0.76. Then specificity, true negative divided by true negative plus false positive. For our problem, it was 0.94. Then sensitivity is true positive divided by true positive plus false negative. For our problem, sensitivity is 0.44.

We got this specificity and sensitivity by taking threshold is 0.5. That is our default value, but the question will come, if it goes above 0.5 or below 0.5, what will happen and what should be the correct value.

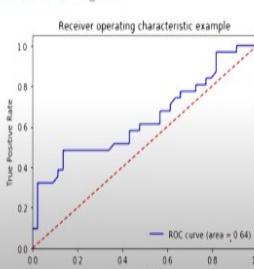
(Refer Slide Time: 02:59)

ROC Curve for Training dataset

```
In [23]: 1 from sklearn.metrics import roc_auc_score
2 from sklearn.metrics import roc_curve, auc
3 log_ROC_AUC1 = roc_auc_score(y_train, y_predict_train)
4 fpr1, tpr1, thresholds1= roc_curve(y_train, y_prob_train)
5 roc_auc1 = auc(fpr1, tpr1)
```



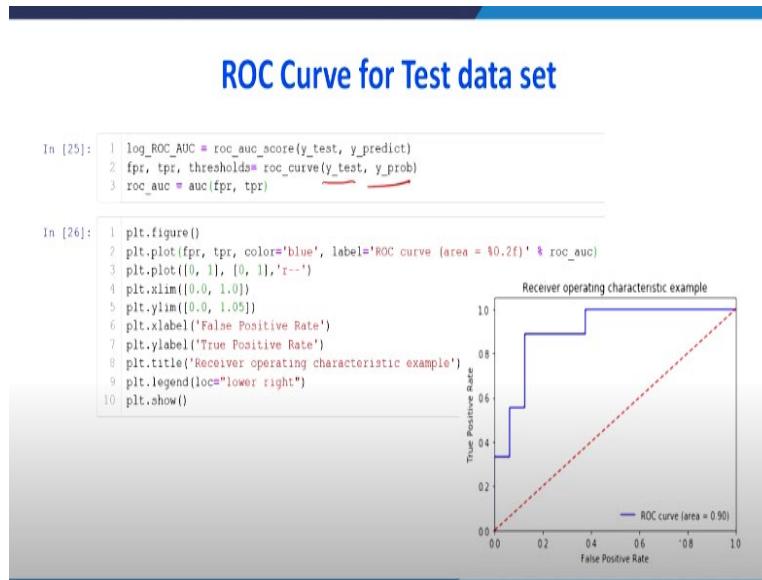
```
In [24]: 1 plt.figure()
2 plt.plot(tpr1, fpr1, color='blue', label='ROC curve (area = 0.2f)'.format(roc_auc1))
3 plt.plot([0, 1], [0, 1], 'r-')
4 plt.xlim([0.0, 1.0])
5 plt.ylim([0.0, 1.05])
6 plt.xlabel('False Positive Rate')
7 plt.ylabel('True Positive Rate')
8 plt.title('Receiver operating characteristic example')
9 plt.legend(loc="lower right")
10 plt.show()
```



Now we will draw the ROC curve for training dataset. So, from sklearn.metrics import roc_auc_score, from sklearn.metrics import roc_curve, auc, log of RUC_AUC1 is equal to

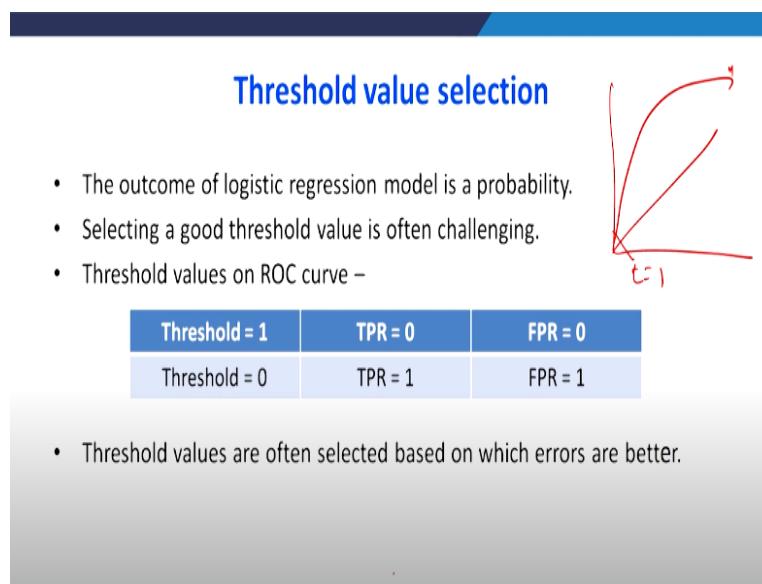
`roc_auc_score`, `y_train`, `y_predict_train`. Then `fpr`, `tpr`. Find out false positive rate and true positive rate and threshold also we are going to test it. So, when you plot it, `fpr` and `tpr`, threshold 1 equal to `roc_curve(y_train, y_prob_train)`, so `roc_auc1` is equal to, so we are going to draw auc curve under `fpr1` is false positive and true positive rate.

(Refer Slide Time: 04:29)



When you plot this, we are getting for different combination of `fpr` and `tpr`, we are getting for default ROC curve. So, the area under AUC for this model is 0.64. Now, let us draw the ROC curve for the Test data set. See only these things are changed, when we draw ROC curve for Test data set, what happened when compared to this, here the AUC, area under curve is increased by 0.9. So what we infer from the previous one, this one is, this model is performing well for the `y test` dataset because the AUC is 0.9.

(Refer Slide Time: 05:03)



How to select threshold value? The outcome of logistic regression model is a probability. Selecting a good threshold value is often challenging. The threshold value on ROC curve, you can take it is 1, if the threshold is 1, what will happen the true positive rate is 0 and false positive rate also 0. This situation. So, this is the place where T equal to 1. When threshold is 0, so true positive rate is 1, this point and false positive rate also 1. So, threshold values are often selected based on which error is better.

(Refer Slide Time: 05:51)

Accuracy checking for different threshold values

At present, randomly we choose some threshold value 0.35. Let us compare by changing different threshold value and verify which threshold value is right choice for us. For that, `y_predict_class1` equal to binarize `y_prob.reshape(1, -1)` by taking 0.35. So, if you take 0.35, this was our predicted values, but if we want to get to the integer value to use `y_predict_class1.astype(int)`, you are getting the integer value and from that, you are getting the confusion matrix.

So, what it says that by keeping threshold values 0.35, so true positive is 9 and true negative is 8. So, we can change different value, for example, for that value, we can check our recall and precision. Now, by taking threshold value as 0.5, let us verify what has happened that confusion matrix value is increased, so the true negative is 15, true positive is 4.

(Refer Slide Time: 07:14)

Accuracy checking for different threshold values

```
In [35]: 1 from sklearn.preprocessing import binarize
2 y_predict_class3 = binarize(y_prob.reshape(1,-1), 0.7)[0]
3 y_predict_class3

Out[35]: array([0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0.])

In [36]: 1 confusion_matrix_4 = confusion_matrix(y_test, y_predict_class3)
2 print(confusion_matrix_4)

[[16  0]
 [ 9  0]]

In [37]: 1 from sklearn.metrics import classification_report
2 print(classification_report(y_test, y_predict_class3))

          precision    recall  f1-score   support

           0       0.64      1.00      0.78      16
           1       0.00      0.00      0.00       9

    micro avg       0.64      0.64      0.64      25
  macro avg       0.32      0.50      0.39      25
weighted avg       0.41      0.64      0.50      25
```

So, what has happened, you recall previously it was 0.5, now it is 0.94. That was the value of 0.50 of the threshold value. If you take threshold value 0.7, what is happening here, the true negative value is increased but there is no true positive at all. So, what is happening after certain threshold value, we are not able to get true positive values but the other values are little improved.

(Refer Slide Time: 07:45)

Calculating Optimal Threshold Value

```
In [38]: 1 from sklearn.metrics import roc_curve, auc

In [39]: 1 fpr, tpr, thresholds = roc_curve(y_test, y_prob)
2 roc_auc = auc(fpr, tpr)

In [40]: 1 print("Area under the ROC curve : %f" % roc_auc)

In [41]: 1 import numpy as np
2 i = np.arange(len(tpr)) # index for df
3 roc = pd.DataFrame({'fpr' : pd.Series(fpr, index=i), 'tpr' : pd.Series(tpr, index = i),
4                      '1-fpr' : pd.Series(1-fpr, index = i), 'tf' : pd.Series(tpr - (1-fpr), index = i),
5                      'thresholds' : pd.Series(thresholds, index = i)})
6 roc.iloc[(roc.tf==0).abs().argsort()[:1]]
```

	fpr	tpr	1-fpr	tf	thresholds
7	0.125	0.888889	0.875	0.013889	0.457033

But the question will come how to choose the right threshold value. We have seen 0.3, 0.5, and 0.7. The appropriate T value that we can get by looking at this table. For example, the output of this program says that the threshold value should be 0.457 is most appropriate for us. Now, Optimal Threshold Value in ROC curve. What has happened here in x-axis, you take 1 minus false positive rate, in y-axis, you take true positive rate. So, at the intersection

point where the true positive rate and true negative rates are intersecting, so that point is considered as the optimal value of threshold value.

(Refer Slide Time: 08:36)

Classification Report using Optimal Threshold Value

```
In [43]: 1 from sklearn.preprocessing import binarize
2 y_predict_class4 = binarize(y_prob.reshape(1,-1), 0.45)[0]
3 y_predict_class4
Out[43]: array([1., 1., 1., 0., 0., 0., 1., 0., 0., 0., 0., 1., 0., 0., 0., 0., 1.,
   1., 0., 1., 1., 0., 0.])

In [44]: 1 confusion_matrix_5 = confusion_matrix(y_test, y_predict_class4)
2 print(confusion_matrix_5)
[[14  2]
 [ 1  8]]

In [45]: 1 from sklearn.metrics import classification_report
2 print(classification_report(y_test, y_predict_class4))
      precision    recall  f1-score   support
          0       0.93      0.88      0.90      16
          1       0.80      0.89      0.84       9

                                ...
   micro avg       0.88      0.88      0.88      25
   macro avg       0.87      0.88      0.87      25
weighted avg       0.89      0.88      0.88      25
```

So, here Classification Report using Optimal Threshold Value, this was our program output. So, here I use binarize y_prob. reshape, so this is our predicted value. So, we got confusion matrix, here true negative is 14, true positive is 8. For that we got the Classification Report also. Here it says that specificity and sensitivity, both are little higher. So, it shows that this is the optimal threshold value.

(Refer Slide Time: 09:13)

The screenshot shows a Jupyter Notebook interface with the following details:

- Title:** Logistic_Regression (unsaved changes)
- Toolbar:** Apps, Google, YouTube, Maps, News, Gmail, Student & Staff Book, Status & Set Book, Logout
- Kernel:** Python 3.0
- Code Cells:**
 - In [17]: `y_prob_train = lreg.predict_proba(x_train)[:,1]`
 - Out[17]: Output is a large array of numerical values ranging from approximately 0.4 to 0.9.
 - In [18]: `y_prob = lreg.predict_proba(x_test)[:,1]`
 - In [19]: `y_prob.reshape(1,-1)`

Now, with help of Python I am going to run the code. I am going to explain how to choose the correct threshold value. So, important necessary library, imported pandas, and imported matplotlib.pyplot then this was the dataset. This dataset I have already discussed with you.

There are two independent variable, and one depend variable, that is Coupon. So, we will describe this dataset. It will give a basic statistics of all the columns.

For spending, there are 100 dataset. The mean is 3.3, standard deviation is 1.74, minimum is 1. So, 25th percentile, 50th percentile, and 75th percentile, maximum is 7.07. For Card, we can look at only the count value, because there is no meaning for mean and standard deviation. Similarly, for Coupon, because both the variables are binary variables. Now, we look at what is the value in the Column coupon.

There are two values there one is 0 and 1, 0 mean that customer did not use the coupon, and 1 means that the customer has coupon. Now, let us see how many 0s and how many 1s by using value_counts function. So, there are 60 people did not use the Coupon and 40 people used the Coupon. So, the baseline method 0.6. Now, we will go for building the LogisticRegression model. I have imported linear_model, sklearn.model_selection.

I have imported train_test_split, there also I have imported LogisticRegression. Then, we will split the dataset by the ratio of 25 percentage dataset for the training, and the remaining 25 percent dataset is for testing. So, let us see how much training dataset, how much test dataset. So, for x variable, the training dataset is 75, for y variable, the training data is 75. For test dataset, x is 25, the test dataset for y is 25. Now, we will construct a LogisticRegression. So, we use the solver lbfgs.

Then, we predict our constructor model with the help of test dataset. In our model, after substituting x values, this was our predicted y value. We can get to know there are different solvers, when you use the LogisticRegression? You can get to know there are different cases. For example, this is help function. You see the multinomial option is supported only by the lbfgs. There are some more method, sag method and newton-cg method that we are not using.

So that is why we have used lbfgs solver for getting this LogisticRegression output. Now, we will predict our model with test data set. This was our predicted y value. Here the input is test data set. Now, we will take the training dataset, then we will predict the model. Because in the training dataset, there are 75 dataset. Here only 25 was there. So, this was our predicted output for the training dataset. Now, we will get the probability value for this.

So, this is the probability for our training dataset. So, there will be 75 dataset is there. There is a different probability. Our question comes what should be the threshold value or the cutoff value. So that we can classify this is 1 or 0. Now for the test dataset also, there should be 25 dataset, we can get the probability. Now, we will run the regression model.

(Refer Slide Time: 13:16)

```

Jupyter Logistic_Regression (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
In [44]: from sklearn.metrics import roc_curve, auc
In [45]: fpr, tpr, thresholds = roc_curve(y_test, y_prob)
roc_auc = auc(fpr, tpr)
In [46]: print("Area under the ROC curve : %f" % roc_auc)
Area under the ROC curve : 0.982778
In [47]: import numpy as np
i = np.arange(len(fpr)) # index for df
roc = pd.DataFrame({'fpr' : pd.Series(fpr, index=i), 'tpr' : pd.Series(tpr, index = i),
                     '1-fpr' : pd.Series(1-fpr, index = i), 'tf' : pd.Series(tpr - (1-fpr), index = i),
                     'thresholds' : pd.Series(thresholds, index = i)})
roc.loc[(roc.tf-0).abs().argmin()][:1]
Out[47]:
          fpr      tpr      1-fpr      tf  thresholds
7  0.125  0.888889  0.875  0.013889  0.457033
In [ ]: fig, ax = plt.subplots()
         plt.plot(roc['tpr'])
         plt.plot(roc['1-fpr'], color = 'red')

```

This is our predicted model, but here we have to show that what should be the threshold value. First, we will check the accuracy of the model. For knowing the accuracy of the model, we have to import accuracy_score. So, the accuracy is 0.76. Next, we can go for constructing a confusion matrix. For that, you have to import a new library called confusion_matrix. When you run this, you are getting confusion matrix.

Here the confusion matrix is see default value is 0.5 is taken, this 15 says true negative and 4 says true positive. The 1 is false positive, the 5 is false negative. So, in our dataset, say that the true negative is 15, false positive is 1, false negative is 5, and true positive is 4. Now, we will get to know what is the Classification Report. In classification report, important things you have to remember. One is the recall; another one is the support.

The Recall gives us an idea about what is actually yes and how often does it predict yes. The Recall value 1 is called as sensitivity, the recall value 0 is called specificity. Precision tells us about when it predicts yes, how often is it correct. So, Precision is true positive divided by true positive by false positive for 1. The accuracy is, the diagonal value of the confusion matrix is the tp and true negative, if you add all the cells value, that is our accuracy.

Now the recall is tp divided by tp plus fn, for value of 1. The f measure is giving the balance between precision and our recall. Now, we will go for finding the accuracy. Accuracy is 0.76. Then, we will go for specificity. The specificity is 0.94, here the default value is 0.5. So, we will go for sensitivity that is true positive rate. For true positive rate, it is 0.44. Now, we will for ROC curve. Now, we will plot that ROC curve.

For default value, the ROC curve is a blue one, which says the area under curve is 0.64. Now, we will see different false positive rate and true positive rate, so this was that one. Now, let us see what are the values of false positive rate. First, we will say fpr. These are the different false positive rate. Now, I will display the output of true positive rate, tpr. So, these values are going to be our x and y axis of our ROC curve.

For, different combinations, we may get different ROC values. So, we will plot ROC curve. This is for our test dataset because you see that here the value which I have taken, this is y set dataset and this is ROC curve. When you look at these two curves, here when you look at this dataset, this is for our training dataset. For training dataset, the ROC curve is like this. The area is 0.64. For the test dataset, the value for ROC curve is 0.9.

So, our model is very well for the test dataset. Now, let us randomly give different T value, different threshold value. Let us see how the ROC curve appears. Suppose we have taken the ROC curve value 0.35, this is threshold value, let us see ROC curve for this value. So, I have predicted values, then value I need in an integer form, I have taken integer form, then I am going to draw the confusion matrix.

So, here the confusion matrix true negative is 8, true positive is 9 because here the value is 0.35, so if it is 0.35, see there are higher true positive rate, that is 9. For this value, let us draw the true positive is 9 because our threshold value is low, everything will be predicted as positive. Now, let us get the Classification Report for that. So, this is our Classification Report. The recall when it is 0, it is 0.5, recall when it is 1, it is 1.00.

Now, let us go for another threshold value that is 0.5, when it is 0.5, see that I have changed the value is 0.5, let us predicted y value. This predicted y value, then let us draw the confusion matrix. What has happened, the threshold value has move on right hand side, we

are getting more true negative and less true positive value. For this, let us get the Classification Report. This 0 represent specificity is increasing.

So what has happening here, sensitivity is decreasing. When you move towards right hand side, the threshold value goes towards right hand side, what is happening specificity is increasing, sensitivity is decreasing. The previous curve you see that, when it is low this side, you see that the sensitivity is 1 almost. It is 1 exactly. When we have low value of, low value is 0.35. When it is 0.35, you see the specificity is 0.5, but sensitivity is 0.1.

When the threshold value is increasing, what has happened that the specificity increased, but the sensitivity decreased. Let us go for 0.7. When the threshold value is 0.7, this is our predicted value. Let us go for confusion matrix. When the threshold value is high, there are more true negative because it is extremely the right hand side. So what they say, whatever the kind of pathology lab that whoever goes there, they will get a negative report.

So, that is the effect of changing the threshold value from lower side to upper side. Now, let us get the Classification Report for this. Here the specificity is 1 when you are having higher threshold value. So, sensitivity is 0 because you have chosen higher threshold value. Now, the important task in our class, we have to choose what is the optimal cut-off point or cut-off point in the sense, optimal threshold value.

So, we will import this ROC curve, then we will run ROC curve, `y_test` and `y_prob`, then we will print area under ROC curve, that is AUC curve, the area under ROC curve for optimal threshold value is 0.90, so it is the best one, because it is nearer to 1. But we want to know what is the optimal threshold value. For that purpose, we have to run this one. We have to import numpy as np, i equal to np.arange for tpr.

For each tpr value, we have to get roc, roc equal to pd. DataFrame, pd. Series false positive rate, then we are getting different index values, so when you run this command, you are getting a table, which shows the optimal threshold value. So, what is the meaning of this one is, if you take the t values 0.457, that will give you higher area under curve.

(Refer Slide Time: 22:15)

```

jupyter Logistic_Regression (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
In [49]: M from sklearn.preprocessing import binarize
y_predict_class4 = binarize(y_prob.reshape(1,-1), 0.45)[0]
y_predict_class4
Out[49]: array([1., 1., 1., 0., 0., 0., 1., 0., 0., 0., 1., 0., 0., 0., 1.,
       1., 0., 1., 0., 1., 1., 0., 0.])

In [50]: M confusion_matrix_5 = confusion_matrix(y_test, y_predict_class4)
print(confusion_matrix_5)

[[14  2]
 [ 1  8]]

In [51]: M from sklearn.metrics import classification_report
print(classification_report(y_test, y_predict_class4))

precision    recall  f1-score   support

          0       0.93      0.88      0.90       16
          1       0.80      0.89      0.84        9

   macro avg       0.88      0.88      0.88       25
weighted avg       0.89      0.88      0.88       25

```

So, when you run this, you see that the blue line says that true positive rate, the red one shows the 1 minus false positive rate. So, this intersection where you see what happening, the true positive rate high, here 1 minus false positive rate also high. So, this is our optimal value. For that optimal value, let us draw our new ROC curve. Here, I have taken 0.45, I have drawn the confusion matrix, you see that here.

Here the confusion matrix 14, the true negative is very high, true positive also very high. So, when you take threshold values 0.45, you are getting higher true negative value and higher true positive value. When you look at Classification Report, you see the specificity value is 0.88, the sensitivity value is 0.89. In this lecture, I have taken a sample problem. With the help of sample problem, I have explained to you how to construct a confusion matrix and how to choose the correct T value, correct threshold value.

We have chosen different threshold value, for example, we have taken threshold value 0.35, we plotted the ROC curve. Then we have taken threshold value 0.5, then we plotted the ROC curve. Next, we have taken threshold value above 0.5 that is 0.7, then we have plotted different threshold value. Then when compared, when we improved or when we increase the threshold value, how the ROC curve differs.

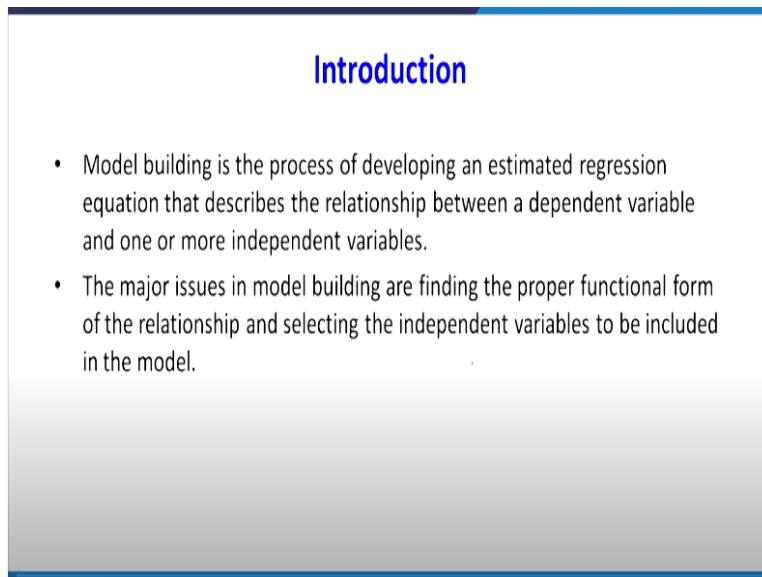
At the end, we have chosen the optimal threshold value. In this problem, we got it as 0.45. Then for that optimal threshold value, we found the AUC, the area under curve, that also very high. So, this is the way to choose the correct threshold value for checking the quality of our classification matrix or Regression models. Thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 44
Regression Analysis Model Building – 1

We have seen so far a simple linear regression model and multiple linear regression model. In this class, we are going to see how to construct a regression model by considering different independent variables, that is model building using regression analysis.

(Refer Slide Time: 00:41)



Introduction

- Model building is the process of developing an estimated regression equation that describes the relationship between a dependent variable and one or more independent variables.
- The major issues in model building are finding the proper functional form of the relationship and selecting the independent variables to be included in the model.

What we are going to do. Model building is the process of developing an estimated regression equation that describes the relationship between a dependent variable and one or more independent variables. What is the meaning of model building? There are different independent variable is there and there is some dependent variable. We are going to find out how to construct a regression model by considering all the independent variable. Whether we have to consider all independent variables or which variable has to be dropped or which variable has to be added.

The major issues in model building are finding the proper functional form of the relationship and selecting the independent variables to include the model. Two concept is there. One is what kind of relationship is going to be found. One is whether it is linear or nonlinear and the other one is how to select the appropriate independent variables. How to select the appropriate independent variables.

(Refer Slide Time: 01:40)

General Linear Regression Model

- Suppose we collected data for one dependent variable y and k independent variables x_1, x_2, \dots, x_k .
- Objective is to use these data to develop an estimated regression equation that provides the best relationship between the dependent and independent variables.

GENERAL LINEAR MODEL

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \dots + \beta_p z_p + \epsilon$$

- z_j (where $j = 1, 2, \dots, p$) is a function of x_1, x_2, \dots, x_k (the variables for which data are collected).
- In some cases, each z_j may be a function of only one x variable.

Suppose, we collect data for one dependent variable y and k independent variables. The independent variables x_1, x_2 and so on and x_k . Objective is to use these data to develop an estimated regression equation that provides the best relationship between the dependent and independent variables. So general form of linear regression model is $y = \beta_0 + \beta_1 z_1 + \beta_2 z_2$ plus and so on plus $\beta_p z_p$ plus error term. Here, the z_j , $j = 1, 2$ up to p is a function of x_1, x_2 , the variables for which the data are collected. In some cases, Z_j may be a function of only one x variable, only one independent variable.

(Refer Slide Time: 02:37)

Simple first-order model with one predictor variable

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

If it is only one independent variable, that is called a simplest first-order model with one predictor variable is this. What is happening here, the value of z is taken only x_1 . There will be an error term. This is the simplest linear regression model.

(Refer Slide Time: 02:52)

Modelling Curvilinear Relationships

- To illustrate, let us consider the problem facing Reynolds, Inc., a manufacturer of industrial scales and laboratory equipment.
- Managers at Reynolds want to investigate the relationship between length of employment of their salespeople and the number of electronic laboratory scales sold.
- Table in the next slide gives the number of scales sold by 15 randomly selected salespeople for the most recent sales period and the number of months each salesperson has been employed by the firm.

Sources: Statistics for Business and Economics, 11th Edition by David R. Anderson (Author), Dennis J. Sweeney (Author), Thomas A. Williams (Author)

Now, we will go for modeling curve linear relationships. This problem is taken from this book, statistics for business and economics, 11th edition by David Anderson, Sweeney and Williams. To illustrate, let us consider the problem facing a company called Reynolds, a manufacturer of industrial scales and laboratory equipment. Managers at Reynolds wants to investigate the relationship between the length of the employment of their salespeople and the number of electronic laboratory scales sold.

So what they want to know, their length of employment of salespeople that is how many years they are working in that company versus number of electronic laboratory scales sold, how much they have sold. Generally, what is the assumption, a person who is having a lot of experience will sell more product. The table in the next slides gives number of scales sold by 15 randomly selected salespeople for the most recent sales period and the number of months each salesperson has been employed by the firm.

(Refer Slide Time: 04:09)

Data

Scales Sold	Months Employed
275	41
296	106
317	76
376	104
162	22
150	12
367	85
308	111
189	40
235	51
83	9
112	12
67	6
325	56
189	19

So, this slide shows the data, scales the product sold, months employed. So, here months employed is going to be our independent variable, scales sold is going to be our dependent variable. This is y and this is our x.

(Refer Slide Time: 04:27)

Importing libraries and table

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm

In [9]: tbl1 = pd.read_excel('Reynolds.xlsx')
tbl1

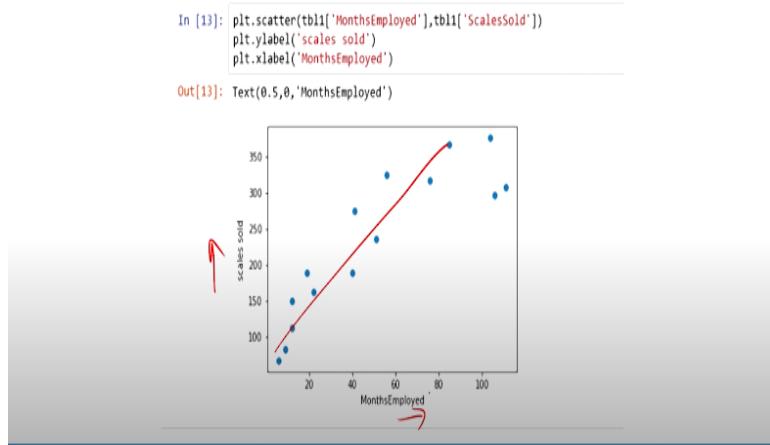
Out[9]:
```

	Scales Sold	MonthsEmployed
0	275	41
1	296	106
2	317	76
3	376	104
4	162	22
5	150	12
6	367	85
7	308	111
8	189	40
9	235	51
10	83	9
11	112	12
12	67	6

With the help of Python, first we will construct a simple linear regression equation; let us see what is happening. I have brought the screenshot of Python programming. At the end of the class, I will run these codes. You can type these commands in your PC and you can verify the answer. Import pandas as pd, import numpy as np, import matplotlib.pyplot as plt, import statsmodels.api as sm. The data, which I have stored in the file name called Reynolds.xlsx. I have read this data. This was the dataset. There is y, this is the x.

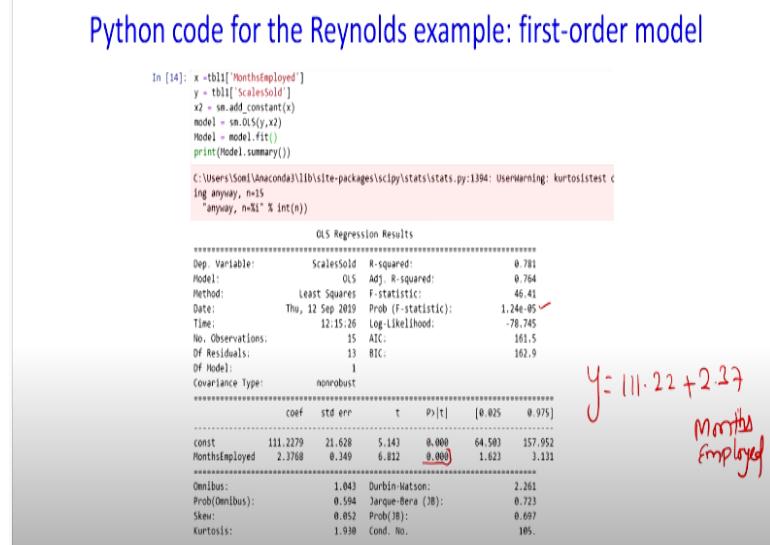
(Refer Slide Time: 05:17)

SCATTER DIAGRAM FOR THE REYNOLDS EXAMPLE



Let us do a Regression Analysis. First, going for Regression Analysis, we will go for a scatter plot. For drawing the scatter plot, plt. scatter tb1, that is my first variable months employed, this variable, second variable is going to be in y-axis scales sold. When it is plotted, it seems to be there is a positive trend. What is the meaning of positive trend? When the number of months employed is increasing and the sales also increasing. It says that the person who is experienced the salesperson can sell more products when compared to an inexperienced person.

(Refer Slide Time: 06:01)



So, this was the code for running Regression. So $x = tb1$, that is months employed is independent variable, $y = tb1['ScalesSold']$ is my dependent variable. I am going to add $x2$ equal to $sm.add_constant$, so that in my Regression model, I will get a constant. So, $model = sm.OLS$. OLS is ordinary least square method, y is dependent variable, $x2$ is independent

variable because in the x2, I am going to have the x variable also. So, Model equal to model.fit(), so print (Model.summary()). So, how to interpret this one.

Look at this constant, say $y = 111.22 + 2.37$ months employed. We will test the significance of the model. First, we will look at the F statistics and the corresponding p value. Here, p value is 1.24 into 10 to the power - 5, very low. As a whole model, this model is significant. Then look at the significant of individual variables. The month employed is independent variable. When we look at the p value, this also less than 5, so we can say the months employed is a significant variable.

(Refer Slide Time: 07:39)

First-order regression equation

$$\text{Sales} = 111 + 2.38 \text{ Months}$$

where

Sales = number of electronic laboratory scales sold

Months = the number of months the salesperson has been employed

This was my model. Here, the sales is that is y variable, number of electronic laboratory scales sold, months equal to number of months the salesperson has been employed.

(Refer Slide Time: 07:51)

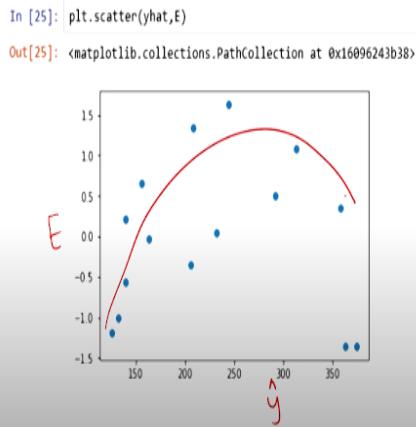
Standardized residual plot for the Reynolds example: first-order model

```
In [18]: E=Model.resid_pearson  
In [19]: E  
Out[19]: array([-1.33945744, -1.35645711,  0.50765089,  0.35510943, -0.03063607,  
   0.20702037,  1.09543558, -1.35411191, -0.34936157,  0.05163116,  
  -1.00000007, -0.50001143, -1.18121025,  1.69231117,  0.65866542])  
  
In [42]: yhat = Model.predict(x2)  
yhat  
Out[42]: 0    288.675693  
1    363.166063  
2    291.862814  
3    358.412511  
4    163.516970  
5    319.749221  
6    313.251788  
7    375.049935  
8    286.299918  
9    232.443442  
10   332.618896  
11   339.749221  
12   125.488571  
13   244.327310  
14   156.386645  
dtype: float64
```

What will happen? First we will plot, the residual plot because as I told you in my previous classes, it is not only the R square fp value and individual significance value is important, the same time, we have to check the residual of that Regression Model. First, we will find the residual $E = \text{Model. resid_pearson}$. So, this was my residual. So, for the $x2$ that is my independent variable, I have predicted the y hat value. This is my predicted y value.

(Refer Slide Time: 08:34)

Standardized residual plot for the Reynolds example: first-order model



Now, I am going to make a plot between in x-axis, we have taken y hat, in y-axis, this is the error. When you look at this picture, in x-axis, it is our y hat. In y-axis, standardized residual. When I look at this standardized residual, it is not coming in the rectangular shape. You see that, there is a possibility of certain kind of curvilinear relationship. You may not agree it is exactly a curve linear relationship because the number of dataset is less. If there are more

number of dataset, we can exactly say. So, what is happening, it is not in the rectangular shape, it is suggesting that there may be curvilinear relationship between x and y.

(Refer Slide Time: 09:25)

Need for curvilinear relationship

- Although the computer output shows that the relationship is significant (p -value .000) and that a linear relationship explains a high percentage of the variability in sales (R-sq 78.1%), the standardized residual plot suggests that a curvilinear relationship is needed.



Very important point, why we have to go for curvilinear relationship. Although the computer output shows that the relationship is significant, because the p value is less than 0.05 and that the linear relationship explains the higher percentage of variability, that is R-square, so R-square is 78.1. The standardized residual plot suggest that the curvilinear relationship is needed.

So, what is the point which I wanted to say, not only R-square, not only the significant value. Apart from that, we have to draw the different residual plot to verify whether the model is correct or not. When we are plotting the standardized residual model, it is suggesting for a nonlinear relationship.

(Refer Slide Time: 10:19)

Second-order model with one predictor variable

- Set $Z_1 = x_1$ and $Z_2 = X^2$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon$$



So, we are going for what kind of nonlinear relationship we are going to have it. Z_1 is x_1 , no problem. Then Z_2 , I am going to square that X value. So, this squared value, that is x_1 squared, is taken as a new independent variable. Previously only x_1 was there, this square of x_1 is a new independent variable. So, this is a general linear model, there is independent variables are having some non-linear patterns. That is x_1 squared.

(Refer Slide Time: 10:54)

New Data set

- The data for the MonthsSq independent variable is obtained by squaring the values of Months.

```
In [29]: X_sq = (x**2)
X_sq
Out[29]: 0    1681
1    11236
2     5776
3    18816
4     484
5     144
6    7225
7   12321
8    1600
9    2601
10    81
11    144
12     36
13    3136
14    361
Name: MonthsEmployed, dtype: int64
```

So, what I am going to do is to prepare a new dataset that is square of x . For that purpose, x_sq , I am naming that way is equal to $x ** 2$, so that is a squared. So, this squared value is going to be taken as another independent variable.

(Refer Slide Time: 11:11)

Python output for the Reynolds example: second-order model

```
In [31]: x_new = np.column_stack((x,X_sq))
x_new2 = sm.add_constant(x_new)
model2 = sm.OLS(y,x_new2)
Model2 = model2.fit()
print(Model2.summary())

OLS Regression Results
-----
Dep. Variable: SalesSold R-squared: 0.903
Model: OLS Adj. R-squared: 0.800
Method: Least Squares F-statistic: 55.36
Date: Thu, 12 Sep 2019 Prob (F-statistic): 8.75e-07
Time: 12:38:01 Log-likelihood: -72.704
No. Observations: 15 AIC: 151.4
Df Residuals: 12 BIC: 153.5
Df Model: 2
Covariance Type: nonrobust
-----
            coef  std err      t  P>|t|   [0.025   0.975]
-----
const    45.3476  22.775   1.991  0.070  -4.274  94.969
x1       6.3448  1.058   5.998  0.000  *** 4.040  8.650
x2      -0.0345  0.009  -3.854  0.002  *** -0.054  -0.015
-----
Omnibus: 2.162 Durbin-Watson: 1.713
Prob(Omnibus): 0.339 Jarque-Bera (JB): 1.093
Skew: -0.126 Prob(JB): 0.696
Kurtosis: 1.758 Cond. No. 1.48e+04
-----
```

You see that `x_new` is `np.column_stack`, see this `x squared`. I wanted to have the constant, so `x_new2` equal to `sm.add_constant x, new`. So, the model 2 equal to `sm.OLS y, x_new2`, so `model.fit`, then `print summary`. So, now what has happened. Look at this point, so there are two independent variable, one is `x1` and `x2`. So, one variable is `x1` is the month, the `x2` is the squared value. Look at the R-square. The R-square is previously 0.7, now it is improved.

So, our model is good. Look at the significance of each variable. One is `x1`, another one is squared value of `x1`, so both are significant value. This also less than 0.05, this also less than 0.05. So, what has happened, when you introduce a squared value, then the model is significant. Not only the significance is enough for us to decide, the model is good or not, we should go for error analysis.

(Refer Slide Time: 12:26)

Second-order regression model

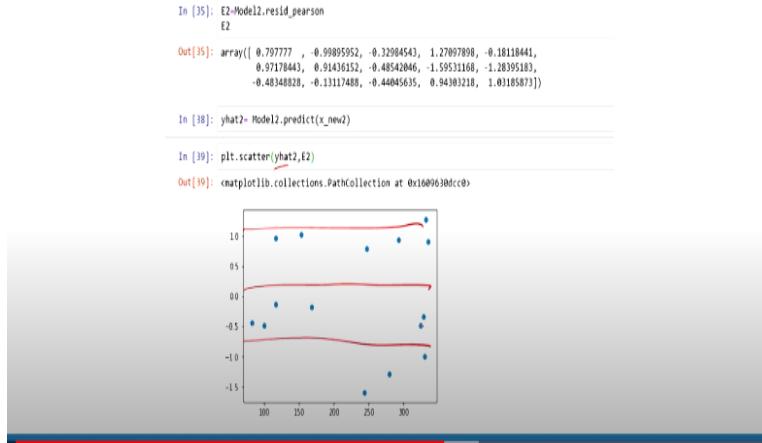
$$\text{Sales} = 45.3 + 6.34 \text{ Months} - .0345 \text{ MonthsSq}$$

MonthsSq = the square of the number of months the salesperson has been employed

So, what is that model which we have created. $45.3 + 6.34$ months, then minus 0.0345. That is squared value of month. Now, look at the standardized residual plot of new variable.

(Refer Slide Time: 12:41)

Standardized residual plot for the Reynolds example: second-order model



Now, look at the standardized residual plot of new variable. So, here what happened, I have found the error term here. In the error term, for the model which you have created. So, I have predicted y hat. So, now I am drawing a graph of standardized residual. So what is happening, it is kind of a linear residual relationship but not only that you can see that there is a possibility of getting a rectangular shape. So that we can say our model is improved.

(Refer Slide Time: 13:15)

Interpretation second order model

- Figure corresponding standardized residual plot shows that the previous curvilinear pattern has been removed.
- At the .05 level of significance, the computer output shows that the overall model is significant (p -value for the F test is 0.000)
- Note also that the p -value corresponding to the t -ratio for MonthsSq (p -value .002) is less than .05
- Hence we can conclude that adding MonthsSq to the model involving Months is significant.
- With an R -sq(adj) value of 88.6%, we should be pleased with the fit provided by this estimated regression equation.

How to interpret the second order model? The figure corresponding to standardized residual plot shows that the previous curvilinear pattern has been removed at 0.05 level of significance, our Python output shows that the overall model is significant because the p

value for the f-test is 0.000. Note also that the p value corresponding to the t-ratio of Months Sq is 0.002. This is also significant. Hence, we can conclude that adding months square as a new variable to the model is significant. With an adjusted R-sq of 88.6%, we should be pleased with the fit provided by this estimated regression equation where there is a nonlinear relationship.

(Refer Slide Time: 14:22)

Meaning of linearity in GLM

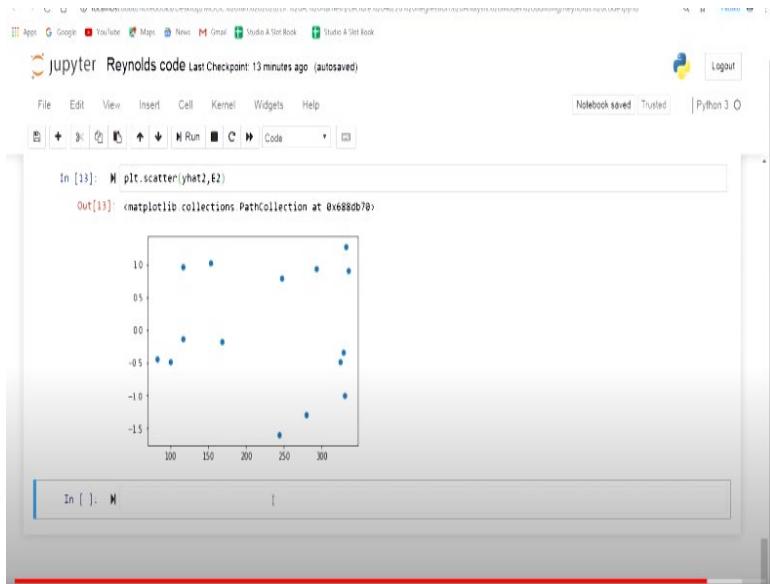
- In multiple regression analysis the word linear in the term "general linear model" refers only to the fact that $\beta_0, \beta_1, \dots, \beta_p$ all have exponents of 1.
- It does not imply that the relationship between y and the x 's is linear.
- Indeed, we have seen one example of how equation general linear model can be used to model a curvilinear relationship.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad y = b_0 + b_1 x_1 + b_2 x_1^2$$

Meaning of linearity in a general linear regression model. In multiple regression analysis, the word linear in the term general linear model refers to the fact that beta0, beta1, up to beta p all have exponents of 1. What is the meaning of this one is, see suppose a model is there, beta0+ beta 1x1 + beta 2x2. When I say linear model, this coefficient of beta1, beta2, beta0, these are linear. There is a x1, y relationship. We are not discussing about relation x and y.

When we say linear, the coefficient beta0, beta1, beta2 are linear. It does not imply the relation between y and x is linear. Indeed, we have seen one example, how linear equation general linear model can be used to model a curvilinear relationship. Previously, we have done one model $y = b_0 + b_1 x_1 + b_2 x_1^2$. Actually, this x_1^2 is nonlinear but we have called it as linear model because this b_0, b_1 and b_2 is linear. Now, I will run the Python code for this model which I have explained.

(Refer Slide Time: 15:55)



Now, in the Python environment, I will tell you how to do the curvilinear relationship. First, I have imported the necessary libraries pandas, numpy, matplotlib, statsmodels, by running this then I am going to import the data. The data which I have stored in the excel file. The file name is Reynolds. This is the data. This data shows the MonthsEmployed independent variable, ScalesSold is our dependent variable.

First, we will go for a scatter plot between these two variable. Now, this scatter plot shows that there is a positive relationship between the MonthsEmployed and the ScalesSold. This implies a person having more experience can sold more products. Now, we will go for a simple linear Regression equation. X equal to MonthsEmployed tbl1, y is our dependent variable tbl1 scales sold, $x_2 = \text{sm.add_constant } x$. Thus I need to have a constant and model = sm.OLS, so model.fit and printing the summary.

So, it shows that when you look at that first one is the R-squared, it is equal to 0.781. The fp value is 1 into 10 to the power - 5, it is very low, so that overall model is significant. Now, look at the independent variable that is our month employed and the corresponding p value is 0.00, so the MonthsEmployed independent variable is also significant independent variable. Now, we can construct a Regression equation. That is, $y = 111.27 + 2.37 \text{ months employed}$.

Now everything is ok, that is not important. Apart from this, we have to draw the residual plots, we have to look at the behavior of the residual plots. That will say whether or model is correct or not. So, I am plotting the residual, so this is my residual value, then the residual

plot what is going to be there in x-axis, I am going to have the y predicted value, in y-axis we are going to have standardized residual value. This is my \hat{y} , y predicted value.

Now, I am going to draw the scatter plot between \hat{y} and standardized residual. Look at this one, there is a curvilinear relationship. It is not the straight line. It is not coming in the rectangular shape, so it is suggesting that instead of going for linear relationship, you go for non-linear or curvilinear relationship that may be the better data model for the given set. So what we are going to do.

We have one independent variable that we are going to square that. You may ask why we have to square. You can go for cube also, you can go for power 3, power 4, power 5, but at the beginning we start with the power 2. So this was my squared value. Now, this variable is taken as another independent variable. Now, we are having two independent variable, one independent variable is `MonthsEmployed`, another independent variable square of that `MonthsEmployed`.

Now that variable also look at this one. `X_sq` that is taken as another independent variable. Now, we will run the Regression equation. Now, what is happening. When you look at this, the R square is previously 0.7, now it is 0.9, so the R-square value is improved. So the model is a good model. The adjusted R-square also 0.886. So, the model is good. You look at the P-value of F statistics.

When you look at the P value, this is very low, less than 0.05, that means the overall model is significant. Now, look at the two independent variable, one is `x1`, another one is `x2` square. Here, the `x2` is nothing but the square of the first independent variable. When you look at the p-value for the first variable, it is 0.00, for the second variable it is 0.002. That is where the squared term. So both the p values are less than 0.05, we can conclude that both independent variables are significant.

It is not enough to check only the individual significant, overall significant and R square. Apart from this, we have to go for residual plot. So, when you go for residual plot for our second model, so this is the standardized residual, now we will go for predicted value of new y , that is \hat{y}_2 , and now we will plot it. Now, it shows there is no curvilinear relationship.

Now, we can say that in this model, we can go for a simple linear relationship when you go for curvilinear relationship that is the best model for a given data.

In this lecture, we have seen how to do a curvilinear regression model. In our previous lecture, I have explained how to do simple linear regression and multiple linear regression model, but in this lecture I have given you an example when we should go for curvilinear relationship between x and y . We have taken one example, in that example, we have taken our first simple linear relationship between x and y .

Then we look at the residual plot, that residual plot suggested that we should go for non-linear relationship, so we have squared that independent variable. Again, we have constructed a new regression model. In that, we have realized that when we look at the residual plot of the new model, we realize that the curvilinear model is the better model for the given data when compared to simple linear relationship.

In the next class, we will go for interaction, how to do if there is interaction between two independent variable x_1 and x_2 , how to do that kind of regression model that we will see in the next class.

Data Analytics with Python
Prof. Ramesh Anbanandam.
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 45
Regression Analysis Model Building (Interaction) – II

In this lecture, we are going to see if there are two independent variable. If they have some interaction, how to incorporate this effect of interaction onto the dependent variable. Before that I will explain with an example what is interaction, then I will construct regression model for incorporating this regression. At this end, I will use the Python to run this interaction regression model.

(Refer Slide Time: 00:55)

Agenda

- Incorporating Interaction of the independent variable to the regression model
- Python demo

The agenda for this lecture is incorporating interaction among independent variables to the regression model and Python demo.

(Refer Slide Time: 01:07)

Interaction

- If the original data set consists of observations for y and two independent variables x_1 and x_2 , we can develop a second-order model with two predictor variables by setting $z_1 = x_1$, $z_2 = x_2$, $z_3 = x_1^2$, $z_4 = x_2^2$, and $z_5 = x_1x_2$ in the general linear model of equation
- The model obtained is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon$$

- In this second-order model, the variable $z_5 = x_1 x_2$ is added to account for the potential effects of the two variables acting together.
- This type of effect is called **interaction**.

First, we will see what is interaction. If the original dataset consist observation for y and two independent variable x_1 and x_2 , we can develop a second-order model with two predictor variables setting $z_1 = x_1$, $z_2=x_2$, $z_3=x_1$ square, and z_4 is x_2 square and z_5 is the fifth independent variable that is x_1 and x_2 in the general linear model equation. So, when you bring this interaction, our regression equation will become like this, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \text{square of the first independent variable } x_1 \text{ square plus the square of the second independent variable } x_2 \text{ square and interaction}$.

In this second-order model, this is called as second order regression model, the variable z_5 , that is x_1 and x_2 is added to the account for the potential effect of two variables acting together. This type of effect is called interaction. So this term we say as interaction.

(Refer Slide Time: 02:19)

Example – Interaction

- A company introduces a new shampoo product.
- Two factors believed to have the most influence on sales are **unit selling price** and **advertising expenditure**.
- To investigate the effects of these two variables on sales, prices of \$2.00, \$2.50, and \$3.00 were paired with advertising expenditures of \$50,000 and \$100,000 in 24 test markets.

Source: Statistics for Business and Economics, 11th Edition by David R. Anderson (Author), Dennis J. Sweeney (Author), Thomas A. Williams (Author)

We take one example problem to understand how to do interaction into the regression model. This problem is taken from this book statistics for business and economics, 11th edition. A company produces a new shampoo product, two factors believed to have the most influence on sales are unit selling price and advertising expenditure. So, there are two variables that is going to affect our sales. One variable is unit selling price, another variable is advertising expenditure.

These two variables are independent variable. The sales is the dependent variable. So, in this problem setting, there is one dependent variable, two independent variables. To investigate, the effect of these two variables on the sales, the prize of 2.5 dollar and 3 dollar were paired with advertising expenditure of 50,000 dollars and 100,000 dollars in 24 test markets. I will show you this dataset.

(Refer Slide Time: 03:24)

Price	Advertising Expenditure (\$1000s)	Sales (1000s)
2	50	478
2.5	50	373
3	50	335
2	50	473
2.5	50	358
3	50	329
2	50	456
2.5	50	360
3	50	322
2	50	437
2.5	50	365
3	50	342
2	100	810
2.5	100	653
3	100	345
2	100	832
2.5	100	641
3	100	372
2	100	800
2.5	100	620
3	100	390
2	100	790
2.5	100	670
3	100	393

This dataset, you say that, there are 3 levels in prize, 2, 2.5, and 3. There are 2 level in the advertising expenditure. One is 50, another one is 100. So that there will be a 24 different alternatives. The last column is sales.

(Refer Slide Time: 03:44)

MEAN UNIT SALES (1000s)

	Price			
	\$2.00	\$2.50	\$3.00	
Advertising Expenditure	\$50,000	461	364	332
	\$100,000	808	646	375
			Mean sales of 808,000 units when price = \$2.00 and advertising expenditure = \$100,000	

Now, we have made a summary of the previous table. What the summary says, when the price is 2 dollars, when the advertising expenditure is 50,000 dollars, this 461 says the mean sales. So, for example, in another sales, look at this one. When the price of the shampoo is 2 dollar, the expenditure is 100,000 dollars, this was the mean of all that combinations. So, the mean sales of 888,000 units when the price is 2 dollars and the advertising expenditure.

How it was done where there is a; 2 is there and you have to look at the corresponding sales value. The average of these four element is our 808. Similarly, the cells is nothing but the mean of that level and the corresponding variable. Similarly, how we got the 461. When the price is 2, advertising expenditure is 50. Because the next one, it is going to the 100. So, the average of this value is 461. Now look at this table.

What it says that by keeping the selling prices to 2 dollars, when you increase the advertising expenditure, the mean value of the sales is increasing. Here it is increasing. The second case by keeping 2005 dollar as the prize, when you increase the expenditure 50,000 dollar to 100,000 dollar what is happening here, your sales is increasing. Here although the sales is increasing. This is one way. By looking at another way, when you find the difference between the 50,000 and 100,000 dollars that we will show you in the next slide, what will happen the difference, instead of increasing, it will start decreasing.

(Refer Slide Time: 06:18)

Interpretation of interaction

- When the price of the product is \$2.50, the difference in mean sales is $646,000 - 364,000 = \underline{282,000 \text{ units}}$.
- Finally, when the price is \$3.00, the difference in mean sales is $375,000 - 332,000 = \underline{43,000 \text{ units}}$.
- Clearly, the difference in mean sales between advertising expenditures of \$50,000 and \$100,000 depends on the price of the product.
- In other words, at higher selling prices, the effect of increased advertising expenditure diminishes.
- These observations provide evidence of interaction between the price and advertising expenditure variables.



This is the explanation for our previous slide. When the price of the product is 2.5 dollars, the difference in mean sale is when it is 2.5 dollar, so the difference in mean sale is 646,000 dollar minus 364,000 dollars, this was your 282,000 dollars, for 3 dollars, the difference in mean sale is 43. So what is happening, the difference in mean sale is decreasing. Clearly, the difference in mean sales between advertising expenditures of 50,000 dollars and 100,000 dollars depends on the price of the product.

In other words, at higher selling prices, the effect of increased advertising expenditure diminishes. Actually what it has to do, when the price of the product increases, then we go for increasing the advertising expenditure, the sales also has to increase, but it is not happening so. So, what is happening when the selling price is increasing, the effect of advertising expenditure on the sale diminishes. These observations provide evidence of interaction between the price and the advertising expenditure variables.

(Refer Slide Time: 07:15)

Interpretation of interaction

- Note that the sample mean sales corresponding to a price of \$2.00 and an advertising expenditure of \$50,000 is 461,000, and the sample mean sales corresponding to a price of \$2.00 and an advertising expenditure of \$100,000 is 808,000.
- Hence, with price held constant at \$2.00, the difference in mean sales between advertising expenditures of \$50,000 and \$100,000 is $808,000 - 461,000 = \underline{347,000 \text{ units}}$.

I am going to interpret this mean unit sale against advertising expenditure. Note that the sample mean sales corresponding to the price of 2 dollars and an advertising expenditure of 50,000 dollars is 461,000, and the sample mean sales corresponding to the price of 2 dollars, and the advertising is 808 dollars. I am referring to this 461 and 808. Hence the prize held constant 2 dollars.

The difference in the mean sales between advertising expenditures 50,000 dollars and 100,000 dollars is 808,000 dollars minus 461,000 dollars, the difference is 347,000. We will go to the next column.

(Refer Slide Time: 08:03)

Interpretation of interaction

- When the price of the product is \$2.50, the difference in mean sales is $646,000 - 364,000 = \underline{282,000 \text{ units}}$.
- Finally, when the price is \$3.00, the difference in mean sales is $375,000 - 332,000 = \underline{43,000 \text{ units}}$.
- Clearly, the difference in mean sales between advertising expenditures of \$50,000 and \$100,000 depends on the price of the product.
- In other words, at higher selling prices, the effect of increased advertising expenditure diminishes.
- These observations provide evidence of interaction between the price and advertising expenditure variables.

When the price of the product is kept 2.50 dollars, the difference in mean sale is 282,000 units. Finally, when the price is 3 dollars, the difference in mean sale is 43,000 units. Clearly,

the difference in mean sales between the advertising expenditure of 50,000 dollars and 100,000 dollars depends on the price of the product. In other words, at higher selling prices, the effect of increased advertising expenditure diminishes.

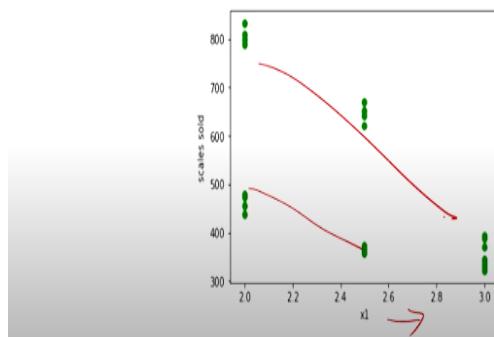
What it happens, when the price increases, when the advertising expenditure also increases, the sales has to increase, but instead of increasing, it starts decreasing. So, the expenditure diminishes. These observations provide evidence of interaction between the price and advertising expenditure variables.

(Refer Slide Time: 08:57)

Mean unit sales (1000s) as a function of selling price

```
In [7]: plt.scatter(tbl1['Price'],tbl1['Sales(1000s)'], color='green')
plt.ylabel('sales sold')
plt.xlabel('x1')
```

```
Out[7]: Text(0.5,0,'x1')
```



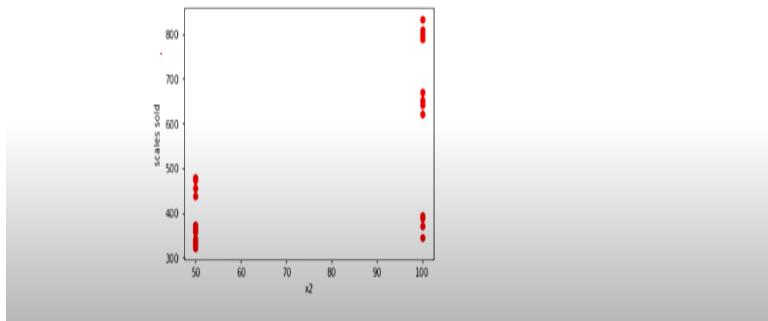
First, we will do the Python code for that. I have imported the data, the prices 2, 2.5, 3 level, there is advertising expenditure is 50 and 100. The sale is in terms of unit, that is 478 and so on. When you plot this scatterplot, see that there are three different levels. What it says that, whenever the price of the product is increasing, the sales it is not increasing. The sales you see that there is a decreasing trend. It has to increase. Why it is decreasing, so there is no effect of amount spent on expenditure when x1 increases.

(Refer Slide Time: 09:47)

Mean unit sales (1000s) as a function of Advertising Expenditure(\$1000s)

```
In [6]: plt.scatter(tbl1['AdvertisingExpenditure($1000s)'),tbl1['Sales(1000s)'], color='red')
plt.ylabel('scales sold')
plt.xlabel('x2')
```

Out[6]: Text(0.5,0,'x2')



So, this graph shows that there is effect of interaction. So this scatterplot shows between the advertising expenditure, there are two level, one is 50,000 dollars, another one is 100,000 dollars. The y-axis is the number of scales sold.

(Refer Slide Time: 10:00)

Need for study the interaction between variable

- When interaction between two variables is present, we cannot study the effect of one variable on the response y independently of the other variable.
- In other words, meaningful conclusions can be developed only if we consider the joint effect that both variables have on the response.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

y = unit sales (1000s)

x_1 = price (\$)

x_2 = advertising expenditure (\$1000s)

In our summary table and our scatterplot, we have realized there is interaction between x_1 and x_2 . When interaction between two variables are present, we cannot study the effect of one variable on the response variable y independently of each variable. In other words, a meaningful conclusion can be developed only if we consider the joint effect of both the variables having the response.

So, what is the joint effect is this x_1 and x_2 . We have realized in that summary table, that there is a interaction between both the variable x_1 and x_2 . Here y is the unit sales, in terms of units, x_1 is the price, it has three level. x_2 is advertising expenditure, it has two levels.

(Refer Slide Time: 10:54)

Estimated regression equation, a general linear model involving three independent variables (z_1 , z_2 , and z_3)

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \epsilon$$

$$z_1 = x_1$$

$$z_2 = x_2$$

$$z_3 = x_1 x_2$$

Now, the estimated regression equation, a general linear model involving 3 independent variables, that is z_1 , z_2 , and z_3 . Here, the z_1 is x_1 , z_2 is x_2 , and z_3 is this interaction variable, that is x_1 multiplied by x_2 . What we have to do, apart from x_1 and x_2 , we have to introduce another variable, that is product of two variable x_1 and x_2 .

(Refer Slide Time: 11:21)

Interaction variable

- The data for the PriceAdv independent variable is obtained by multiplying each value of Price times the corresponding value of AdvExp.

```
In [11]: z1 =tbl1['AdvertisingExpenditure($1000s)']
z2 =tbl1['Price']
z3 =z1*z2
```

Now, we will create a new variable, that is z_3 , that is the product of z_1 and z_2 . The data for price advertisement of independent variable is obtained by multiplying each value of the

price times, the corresponding value of advertising expenditure. So, both variable z1 and z2 has to be multiplied, that will be our new variable.

(Refer Slide Time: 11:43)

New Model

```
In [12]: x_new = np.column_stack((z1,z2,z3))
y = bdf['Sales(1000s)']
xnew2 = sm.add_constant(x_new)
model2 = sm.OLS(y,xnew2)
Model2 = model2.fit()
print(Model2.summary())

OLS Regression Results
-----
Dep. Variable: Sales(1000s) R-squared: 0.978
Model: OLS Adj. R-squared: 0.975
Method: Least Squares F-statistic: 297.9
Date: Thu, 12 Sep 2019 Prob (F-statistic): 9.26e-17
Time: 13:12:52 (log-likelihood): -111.99
No. Observations: 24 AIC: 232.8
Df Residuals: 20 BIC: 236.7
Df Model: 3
Covariance Type: nonrobust
-----
            coef  std err      t  P>|t|   [0.025  0.975]
-----
const    -275.8333  112.842  -2.444  0.024  -511.218  -40.449
x1       19.6800   1.427  13.788  0.000   16.783   22.657
x2       175.0000  44.547   3.928  0.001   82.077  267.923
x3      -6.08000  0.563  -10.798  0.000  -7.255  -4.905
-----
Omnibus: 0.642 Durbin-Watson: 2.842
Prob(Omnibus): 0.726 Jarque-Bera (JB): 0.565
Skew: 0.335 Prob(JB): 0.754
Kurtosis: 2.661 Cond. No. 4.53e+03
-----
```

After multiplying, now this is our output model for our interaction. So look at the R-square. R-square is 0.978, x1 is our one independent variable, x2 is another independent variable. This x3 is the interaction.

(Refer Slide Time: 12:53)

New Model

$$\text{Sales} = -276 + 175 \text{ Price} + 19.7 \text{ AdvExp} - 6.08 \text{ PriceAdv}$$

where

- Sales = unit sales (1000s)
- Price = price of the product (\$)
- AdvExp = advertising expenditure (\$1000s)
- PriceAdv = interaction term (Price times AdvExp)

So, for this how we can write the regression equation. -276+175 price, that is our x2, then 19.7 advertising expenditure, that is our x1. The third one is our interaction variable, that is x3, that is -6.08. Look at the p value of f statistics, that is very low, the overall model is significant. For all variables, x1, x2 and interaction variables, look at the p-value this one, all are less than 0.05, so each independent variable is significant variables.

(Refer Slide Time: 12:46)

New Model

$$\text{Sales} = -276 + 175 \text{ Price} + 19.7 \text{ AdvExp} - 6.08 \text{ PriceAdv}$$

where

Sales = unit sales (1000s)

Price = price of the product (\$)

AdvExp = advertising expenditure (\$1000s)

PriceAdv = interaction term (Price times AdvExp)

So what is the new model now. Sales equal to $-276+175$ price + 19.7 AdvExp – 6.08 the price end advertisement. This is our interaction term. How to interpret this.

(Refer Slide Time: 13:05)

Interpretation

- Because the model is significant (p -value for the F test is 0.000) and the p -value corresponding to the t test for PriceAdv is 0.000, we conclude that interaction is significant given the linear effect of the price of the product and the advertising expenditure.
- Thus, the regression results show that the effect of advertising expenditure on sales depends on the price.

Because the model is significant, the p -value for the F test is 0.0000 and the p value corresponding to the t test PriceAdv is 0.00, we conclude that interaction is significant given the linear effect of the price of the product and the advertising expenditure. Thus, this regression results shows that the effect of advertising expenditure on sales depends on the price.

(Refer Slide Time: 13:18)

Transformations Involving the Dependent Variable

$$y = b_0 + b_1 x_1 + b_2 x_2$$

$x_2 = 0, 1$

Miles per Gallon	Weight
28.7	2289
29.2	2113
34.2	2180
27.9	2448
33.3	2026
26.4	2702
23.9	2657
30.5	2106
18.1	3226
19.5	3213
14.3	3607
20.9	2888

So far, we have done some transformations only on independent variable. For example, $y = b_0 + b_1 x_1 + b_2 x_2$. Suppose, x_2 is a categorical variable, assume that. What you have done, if x_2 can have only two variables, say 0, 1 gender. So we have done a modification. We have introduced a dummy variable and we have done the model. Now, there may be a situation that your y variable also has to be transformed.

(Refer Slide Time: 14:23)

Model 1

```
In [4]: x = tbl1[['Weight']]
y = tbl1['MilesperGallon']
x2 = sm.add_constant(x)
model = sm.OLS(y,x2)
Model = model.fit()
print(Model.summary())
```

C:\Users\Soni\Anaconda\lib\site-packages\scipy\stats\stats.py:1394: UserWarning: kurtosistest only valid for n >= 20. "anyway, n=%d" % int(n))

Dep. Variable:	MilesperGallon	R-squared:	0.935			
Model:	OLS	Adj. R-squared:	0.929			
Method:	Least Squares	F-statistic:	344.8			
Date:	Thu, 12 Sep 2019	Prob (F-statistic):	2.85e-07			
Time:	15:27:08	log-Likelihood:	-71.891			
No. Observations:	12	AIC:	48.18			
Df Residuals:	10	BIC:	49.15			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	56.0957	2.582	21.725	0.000	50.342	61.849
Weight	-0.0116	0.001	-12.032	0.000	0.014	-0.009

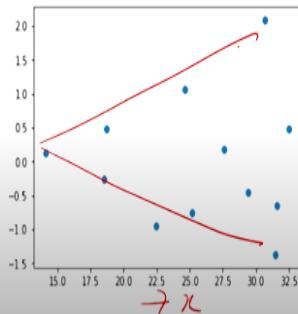
Omnibus: 2.266 Durbin-Watson: 2.213
 Prob(Omnibus): 0.322 Jarque-Bera (JB): 0.952
 Skew: 0.690 Prob(JB): 0.521
 Kurtosis: 3.025 Cond. No.: 1.41e+04

Suppose, the miles per gallon, that is your y variable. This weight is your independent variable. Suppose, if you do a regression analysis for this dataset, see that there is a negative relationship. When the weight increases, the miles per gallon decreases. There is a negative relationship for scatterplot. Now, when you look at the Regression model, the Regression model, y is equal to $56.0957 - 0.0116$. This is significant.

(Refer Slide Time: 14:40)

Standardized residual plot corresponding to the first-order model

```
In [8]: plt.scatter(yhat,E)
Out[8]: <matplotlib.collections.PathCollection at 0x23f77072a58>
```



Now, look at the standard residual plot. First, we will predict the residuals, then we will standardize. Then, we will predict the y value. Now, we will draw a graph between predicted y hat and standardize residual. When we look at this, you see that there is a conical relationship. What is happening, whenever the value of x increases, the variance is not constant. This is violating our model.

What is the model? When the variance or the error term should be same for all value of x, but now what is happening, when the value of x is increasing, the variance also increases. So, it is not fitting to our assumption of regression equation. We are going to take log of y, so the y is there, so we are going to take log of y values = $b_0 + b_1 x_1$. This is going to be the same. Our independent variable will not be disturbing.

But for the dependent variable, we are going to take the log of; the purpose of taking log is that the error term, instead of getting this conical shape, we may get a kind of a rectangular shape. So, that means the variance of the error terms is going to be same.

(Refer Slide Time: 16:02)

Model 2

```
In [12]: Y = np.log(Y)

In [13]: model2 = sm.OLS(Y,x2)
          Model2 = model2.fit()
          print(Model2.summary())

OLS Regression Results
-----
Dep. Variable: MilesperGalon   R-squared:      0.948
Model:           OLS   Adj. R-squared:     0.942
Method: Least Squares   F-statistic:    181.2
Date: Thu, 12 Sep 2019   Prob (F-statistic):  9.84e-08
Time: 15:34:13   Log Likelihood:     17.005
No. Observations: 12   AIC:             -30.01
Df Residuals:    10   BIC:             -29.84
Df Model:        1
Covariance Type: nonrobust
-----
            coef  std err      t      P>|t|      [0.025      0.975]
-----  
const    4.5242   0.099    45.553    0.000    4.303    4.746
Weight   -0.0005  3.72e-05  -13.462   0.000    -0.001   -0.000
-----
Omnibus:       0.099 Durbin-Watson:      2.284
Prob(Omnibus): 0.618 Jarque-Bera (JB):      0.379
Skew:          0.484 Prob(JB):        0.677
Kurtosis:      2.111 Cond. No.       1.43e+04
```

First, what you have done, I have taken log of all dependent variable, that is I call it Y. Now, this log of Y is taken as the new dependent variable. After substituting this, you look at the new variable, one is weight, the R square is increased, and F is good, the model is okay. Now, we will go for the residual plot for this.

(Refer Slide Time: 16:24)

Residual plot for model 2

```
In [14]: E2=Model2.resid_pearson
E2
Out[14]: array([-0.31630114, -1.42005514,  1.5623004,  0.48370101, -0.0537228,
 1.60448776, -0.29474869, -0.79674991, -0.18335787,  0.87474775,
-0.87956572, -0.58073564])

In [15]: yhat = Model2.predict(x2)
yhat
Out[15]: 0    3.377221
1    3.465414
2    3.431840
3    3.297547
4    3.508009
5    3.170268
6    3.192817
7    3.466922
8    2.987694
9    2.914208
10   2.716776
11   3.077064
dtype: float64
```

When you go for residual plot against y hat, this is our standardize residual, so what is happening.

(Refer Slide Time: 16:34)

- The miles-per-gallon estimate is obtained by finding the number whose natural logarithm is 3.2675.
- Using a calculator with an exponential function, or raising e to the power 3.2675, we obtain 26.2 miles per gallon.

$$\text{LogeMPG} = 4.52 - 0.000501 \text{ Weight}$$

$$\text{LogeMPG} = 4.52 - 0.000501(2500) = 3.2675$$

e

Now, there is no conical shape, there is rectangular shape is appearing, but you should be very careful while interpreting the answer because it is not actual y , it is log of y . So, when you substitute the values into this, the miles per gallon estimate is obtained by finding the number whose natural logarithm is 32.675. So what you have to do, suppose if you substitute weight is 2500, we are getting the log of y value, that is miles per gallon is 3.26.

If you want to know the actual value, you take e to the power 3.26, that is why to bring you to normal term, you have to take natural logarithm is 3.26 using a calculator or any exponential function using our Python, we have to rising e to the power 3.26, you will get 26.2 miles per gallon, that is your original y values.

(Refer Slide Time: 17:37)

Nonlinear Models That Are Intrinsically Linear

$$E(y) = \beta_0 \beta_1^x$$

$$E(y) = 500(1.2)^x$$

$$\log E(y) = \log \beta_0 + x \log \beta_1$$

$$y' = \log E(y), \beta'_0 = \log \beta_0, \text{ and } \beta'_1 = \log \beta_1,$$

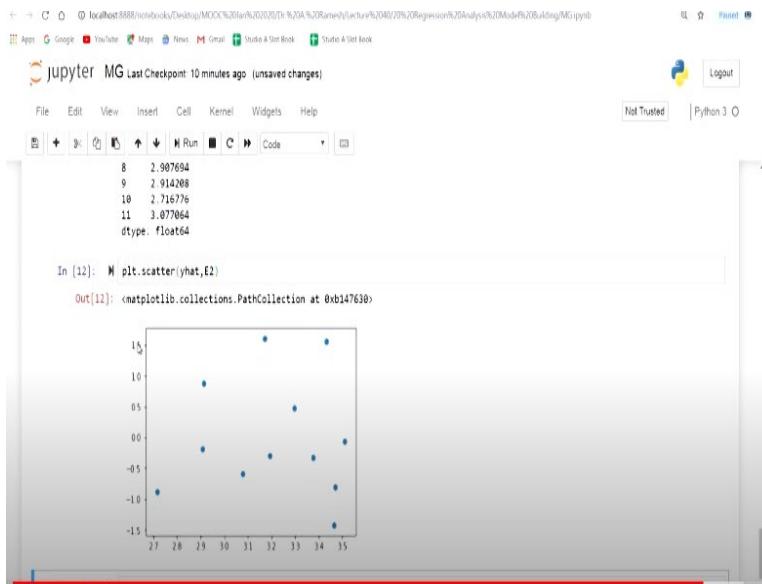
$$y' = \beta'_0 + \beta'_1 x$$

$$\hat{y}' = b'_0 + b'_1 x$$

There are some more nonlinear model. How to do that one, I will explain. Suppose, there may be a nonlinear relationship that the power is there, $\beta_0 + \beta_1 x$, so the expected value, suppose if you substitute β_0 is 500, it is 1.2 to the power x . So for this kind of model, you take log of both the sides. It will become log of expected value of y . So log of $\beta_0 + \log x$ log β_1 . Here the constant term is, this can be written y dash equal to β_0 dash, β_1 dash x .

So, the y dash is the log of $E(y)$, β_0 dash is log of b_0 and β_1 dash, log of β_1 . This equation can be estimated with the sample of this Regression equation, but we should be very careful while interpreting, you have to remember it has to be brought into the original term.

(Refer Slide Time: 18:37)



Now, we are going to do the interaction among independent variable with y with the help of this Python code. So I have imported, the file name is Tyler. So, this was our portion of our file name. First, we should do the scatterplot. So, what is happening here, when the price of the product is increasing, see that the y variable, it is the number of scales sold, it is decreasing. So, this table is suggesting that there is a interaction effect between the prize and the dependent variable.

Look at this. These are another dependent variable, that is advertising expenditure. This also shows that whenever the advertising expenditure is increasing, the car sold is increasing, but it is not linearly increasing because there seems to be some other variable, which is affecting the advertising expenditure. That variable is nothing but the prize. From our scatterplot, plot number one and plot number 2, we realize that there is interaction effect.

So the two variable that is z_1 and z_2 are multiplied. What is our z_1 variable, that is our advertising expenditure, our z_2 variable is price, so new variable is z_1 multiplied by z_2 . We will do this one. Now, the third variable, that is new variable taken as another dependent variable. Now, there are 3 variable, one is for advertising expenditure, another variable for prize, the third one is interaction among these two.

So, when you run this model, we are getting all these three variables, that is x_1 , x_2 , and x_3 is our interaction variable. All are significant. So, we can say that there is interaction effect between x_1 and x_2 . See, our R square is better, 0.978, our fp value also very less, and individual significance of each independent variable is also less than 0.05, also variables are significant.

Now, in our class I have explained one more problem, that is how to do transformation of our dependent variable. So, I have imported the necessary libraries with the data file is this one. So, here the weight is independent variable, but the miles per gallon is dependent variable. So, when you do the scatterplot between these two, there seems to be a negative relationship. When you do a simple linear regression by taking x's weight independent variable, y is the miles per gallon, we are getting this one.

So even though the model is significant when you go for residual plot. What is happening between standardize residual and predicted value? there is a conical shape is there. So, what this implies that the value of x increases, the variance or the error term is not the same. It is getting increased. This is violation of Regression model. To compensate this, we are going to do the transformation, log transformation of our dependent variable. After log transformation when you do again, there is a regression equation.

So, you look at this the third one, now the new dependent variable is the log of y, so the independent variable. So, this one, we will go for a standardized residual plot. Now, what is happening when you go for that, now there is no conical relationship. Then we can say that the log of transformation of dependent variable is correct, you should go for log transformation of our dependent variable.

In this lecture, we have seen how to incorporate if there is interaction among variable, how to incorporate this interaction into our Regression model. We have taken one sample example, when we are plotting the summary table, we have realized that there is a interaction between two variables, then we have taken the product of the two variable that introduces a third variable, then we have done a multiple regression model, we realize that the interaction is significant.

In another problem, what we have seen in this class is, generally we do the transformation in the independent variable, but sometimes, we need to do the transformation for the dependent variable also. So what transformation we have done, we have done log of our y value. Before doing log of y value, we have realized that the variance of the error is not the same. After doing the log transformation, we have realized that the variance of the error term is same, then we have accepted that taking log of our dependent variable is correct. Thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

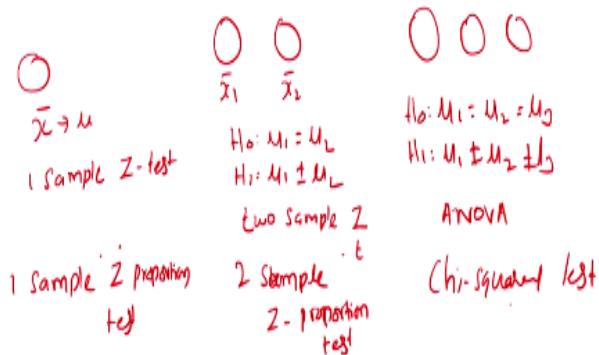
Lecture – 46
Chi-Square Test of Independence-I

Welcome students. Today we are going to see a new topic that is chi-square test. Chi-square test has 2 applications, one is to test the independence, second one is to test the goodness of fit. In this class, we are going to see the test of independence.

(Refer Slide Time: 00:44)

Agenda

- To understand χ^2 Test of Independence



The agenda for this class is to understand chi-square test of independence. Before going to the topic let us see when will you go for chi-square test. In the beginning of the lecture I have explained different types of data nominal, ordinal, interval ratio. Whenever the data is nominal or ordinal you have to go for your test called chi-square test because data which are nominal in nature you cannot go for Z test, you cannot go for T test or ANOVA, even regression it cannot be done.

So now I will explain how this connection has with other test. For example, we might have seen we have studied one sample T test suppose there was a sample 1 was there we have seen with the help of x bar we have predicted μ . After that we have seen 2 sample T test, this is one sample Z test. Whenever there is a 2 population, population 1, population 2. See this is x_1 bar this is x_2 bar.

Here what we have compared what was our null hypothesis $\mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$. Here we are comparing 2 population at a time. So here what we have done we have done 2 sample we write we have done 2 samples Z test similarly we have done 2 sample T test also 1 this is 2. Suppose there are 3 population, population 1 and population 2 and population 3. Here suppose we want to compare the mean of this population.

Our null hypothesis will be $\mu_1 = \mu_2 = \mu_3$; alternative hypothesis is $\mu_1 \neq \mu_2 \neq \mu_3$. Here what we have done whenever we want to compare more than 2 population we have gone for ANOVA right. In the same way you see there is 1 sample here we can say 1 sample proportion test, 1 sample Z proportion test. Suppose if you want to compare the proportion of 2 population we can do 2 sample Z proportion test.

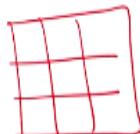
There may be a situation where you have to compare the proportion of more than 2 sample that time we should go for this chi-square test. Suppose if you are comparing mean where the population is more than 2 you should go for ANOVA. Suppose if you are comparing the proportion where you have to compare more than proportion of more than 2 population you should go for chi-square test that is the logic of using chi-square test.

Even if there are 2 sample mean instead of using 2 sample Z test you can use ANOVA also you will get the same result. Similarly, there are 2 sample proportions if you want to compare their proportion instead of using 2 sample proportion test you can do the chi-square test also you will get the same answer because this ANOVA and chi-square test is a generalized format. If you want to compare only 2 that time also you can use ANOVA okay.

(Refer Slide Time: 04:31)

χ^2 Test of Independence

- It is used to analyze the frequencies of two variables with multiple categories to determine whether the two variables are independent.
- Qualitative Variables
- Nominal Data



Now we will go to our topic today's topic that is the Test of Independence. This test is used to analyze the frequencies of two variables with multiple categories to determine whether the two variables are independent or not. So whenever there is a qualitative variables whenever the data is nominal data we should go for test of independence. Here, we are going to have 2 category suppose this is called table contingency table. There will be some value in row, some value in the column. So we are going to see whether these are dependent or independent with the help of an example the next I will explain.

(Refer Slide Time: 05:19)

χ^2 Test of Independence: Investment Example

- In which region of the country do you reside?
A. Northeast B. Midwest C. South D. West
- Which type of financial investment are you most likely to make today?
E. Stocks F. Bonds G. Treasury bills

		Type of financial Investment		
		E	F	G
		n_E	n_F	n_G
Geographic Region	A			O_{13}
	B			
	C			
	D			
				n_A
				n_B
				n_C
				n_D
				N

For example, suppose you have conducted a questionnaire in that questionnaire this is an example of investment example you asked in which region of the country do you reside there was a 4 options. First option is Northeast, Midwest, South, West and you have asked another

question also which type of financial investment are you most likely to make today. The options were stocks, bonds and treasury bills.

This dataset I have captured in the form of a here table so this table is called contingency tables. See in rows I have captured the geographic regions say Northeast, Midwest, South, West. In column I have captured what type of financial investment they are going to make so that I have E, F, G. Suppose I wanted to make an assumption I wanted to test is there any connection between, is there any dependency between the geographic regions where they reside and the type of investment they are willing to make.

So there are 2 variables; one is geographic region another variable is type of financial investment whether these are dependent or independent. So this kind of examples, this kind of problems can be solved with the help of this chi-square test okay. So what is your null hypothesis the geographic regions and the type of investment which they makes are independent there is no connection. Alternative hypothesis it is not independent there is a dependency.

(Refer Slide Time: 07:11)

χ² Test of Independence: Investment Example			
Contingency Table		Type of Financial Investment	
Geographic Region	A	E	F
		e_{12}	
A			
B			
C			
D			
	n_E	n_F	n_G
			n_A
			n_B
			n_C
			n_D
			N

This was a theory behind this test of independence. Suppose if A and F are independent A, F. What is the A here A means the first option that was he belongs to Northeast region, F means he is willing to invest in the bond. If A and B are independent event, we can write P (A intersection F) is P (A) . P (F), then we will find out what is P (A). P(A) is your $n_A / \text{total } N$, capital N that is number of sum of all the elements.

Then $P(F)$ is n_f divided by total element. So when you multiply intersection of P of A and F so that will be n_A divided by N , multiplied by n_F divided by N . Now you will get here in the terms of probability, but if you want to get in terms of frequencies so that has to be multiplied by your capital N . So expected value of $AF = N$ multiplied by $P(A \text{ intersection } F)$.

So it is N . For $P(A)$ we know it is n_A / N , $P(F)$ we know it is n_F / N . So this $N - N$ gets cancelled the remaining is $(n_A \cdot n_F) / N$. So this is very frequent formula which we are going to use. Suppose if you wanted to know this expected values in the cells what you have to do. You have to multiply the row sum and column sum divided by your capital N .

So row sum multiplied by column sum divided by total number elements that will give you the expected value of each cell. One more thing you see that we are multiplying by N here, N is because if you put $P(A \text{ intersection } F)$ we will get only probability if you want to get the answer in terms of frequency that has to be multiplied by your N that is why we are getting this N .

(Refer Slide Time: 09:21)

χ^2 Test of Independence: Formulas

$$E_{ij} = \frac{(n_i)(n_j)}{N}$$

where:

- i = the row
- j = the column
- n_i = the total of row i
- n_j = the total of column j
- N = the total of all frequencies

So expected frequency is E_{ij} is n_i : i represents row j represents the column, n_i means the total of row i multiplied by n_j ; the total of column j divided by capital N the total of all frequencies. This is our expected frequencies. Obviously there is one more frequency which you have to find out the observed frequency that will be given in the problem itself.

(Refer Slide Time: 09:52)

χ^2 Test of Independence: Formulas

Calculated χ^2
(Observed χ^2)

$$\chi^2 = \sum \sum \frac{(f_o - f_e)^2}{f_e}$$

where: $df = (r - 1)(c - 1)$
 $r = \text{the number of rows}$
 $c = \text{the number of columns}$

Then how to find out the test statistics of our chi-square so the calculated chi square our observed chi-square value is the sigma of sigma see first one is observed frequency f_o minus expected frequency whole square / expected frequency. You see that this square is not for this expected frequency that is only for the numerator and another important thing is degrees of freedom.

The degrees of freedom is row – 1, multiplied by column – 1, that means that many number of independent cells we can supply any values. So r represents number of rows c represents number of columns.

(Refer Slide Time: 10:40)

Example for Independence

We will take one example using the theory which I have taught you so far. We will solve a problem of test of independence.

(Refer Slide Time: 10:49)

χ^2 Test of Independence

H_0 : Type of gasoline is independent of income

H_a : Type of gasoline is not independent of income

Suppose before starting any hypothesis testing problem the first step is to first write the null hypothesis. The null hypothesis is the type of gasoline is independent of income. Alternative hypothesis is type of gasoline is not independent of income. Generally, there are different type of gasoline we have a assumption that people were having higher income they will go for good quality in fuel. So we are going to test this is there is any connection, is there is any dependency between their level of income and the type of gasoline they prefer.

(Refer Slide Time: 11:33)

χ^2 Test of Independence

		Type of Gasoline		
		c = 3	Regular	Premium
r = 4		Extra Premium		
Income				
Less than \$30,000				
\$30,000 to \$49,999				
\$50,000 to \$99,000				
At least \$100,000				

This was our problem setup. So in the rows I have captured their level of income less than 30,000 dollar next category is \$30,000 to \$49,999 next is \$50,000 to \$99,000 it is more than \$100,000. In column I have asked what type of gasoline you are using, fuel you are using

regular, premium, extra premium. Generally, there is an assumption whenever the income level is increases the people may go for good quality fuel.

So there is a dependency that is our assumption, there may be a dependency between their level of income and the type of fuel which they use. Here the rows there are 4 rows 1, 2, 3, 4 there are 3 columns so $r = 4$ $c = 3$.

(Refer Slide Time: 12:40)

χ^2 Test of Independence: Gasoline Preference Versus Income Category

$$\begin{aligned}\alpha &= .01 \\ df &= (r - 1)(c - 1) \\ &= (4 - 1)(3 - 1) \\ &= 6 \\ \chi^2_{\text{cal}} &\approx 16.812\end{aligned}$$



If $\chi^2_{\text{cal}} > 16.812$, reject H_0 .
If $\chi^2_{\text{cal}} \leq 16.812$, do not reject H_0 .

First we will find out the degrees of freedom. Assume that the alpha is 1 percentage the degrees of freedom is row - 1 multiplied by column - 1 there are 4 rows $4 - 1$ so $4 - 1$ is 3 there are 3 column $3 - 1 = 2$, so 3 into 2 it is 6. So for 6 degrees of freedom so the chi-square distribution is the right skewed distribution it will be like this. So when right side area is 1 percentage for 6 degrees of freedom the value which got from the table is 16.812. The next slide I will tell you this value we can find out with the help of python.

This was the value which we got from the table. Next what we have to do, we have to calculate the chi square value if our calculated chi-square value how we will calculate it using this formula observed frequency – expected frequency whole square divided by expected frequency. Using this formula, we have to find out the chi-square calculated. If that value is greater than 16.82 we will reject our null hypothesis. If it is less than 16.812 we will accept our null hypothesis.

(Refer Slide Time: 14:06)

Python code

```
In [5]: import pandas  
import numpy  
from scipy import stats
```

```
In [6]: stats.chi2.ppf(0.99,6)
```

Out[6]: 16.811893829770927

As I told you with the help of python import pandas, import numpy, from scipy import stats stats.chi2.ppf when it is a 0.99 because if we want to know one percentage it is 0.99 for 6 degrees of freedom this is 16.811 that sort of value we got it.

(Refer Slide Time: 14:27)

Gasoline Preference Versus Income Category: Observed Frequencies

Income	Type of Gasoline		
	Regular	Premium	Extra Premium
Less than \$30,000	85	16	6
\$30,000 to \$49,999	102	27	13
\$50,000 to \$99,000	36	22	15
At least \$100,000	15	23	25
	238	88	59
			385

This was the data which is given is what is the value which is their inside the cell it is called observed frequency. So what is the meaning of this 85 those were having income less than \$30,000 they have gone for regular type of gasoline. What is the premium, what is the interpretation of this 16 those were having income less than \$30,000 only 16 people have gone for premium type right this one.

You see that when the level of income is increasing you see that here also the number is the people gone for extra premium also increasing. It seems to be that, there is a dependency

between their level of income and type of gasoline they choose. So this is the given data which we have captured. So the first step is we have to find out the row total. The first row total is 107 second row total is 142 third row 73, fourth row 63. Then finding the column total the first column total is 238, second column total is 88, third column total is 59. The value which are given is called observed frequency.

(Refer Slide Time: 15:47)

Gasoline Preference Versus Income Category: Expected Frequencies

		Type of Gasoline			
		Regular	Premium	Extra Premium	
Income	Less than \$30,000	(66.15)	(24.46)	(16.40)	107
	\$30,000 to \$49,999	85	16	6	
\$50,000 to \$99,000	(87.78)	(32.46)	(21.76)	13	142
	102	27		13	
At least \$100,000	(45.13)	(16.69)	(11.19)	15	73
	36	22		15	
		(38.95)	(14.40)	(9.65)	
		15	23	25	63
			238	88	59
					385

Now we should go for our expected frequency here the expected frequency value is given in the bracket. For example, how we got this 66.15 this 66.15 is nothing, but multiplication of row total 107 and the column total 238 divided by 385. So that value is nothing, but 66.15 that is given in the bracket. So the values which are given in the bracket it is called expected frequency.

The value which is not in the bracket it is called observed frequency that is the data which are given to us. So for the second dataset how we have got 24.46 so row total 107 column total 88 / 388 so we will get 24.46. For third one row total 107 column total 59 the total value is 385 like this we have to find out all the cells, all the cells after finding which was given in the bracket. Now we will go for chi-square calculated value.

(Refer Slide Time: 17:01)

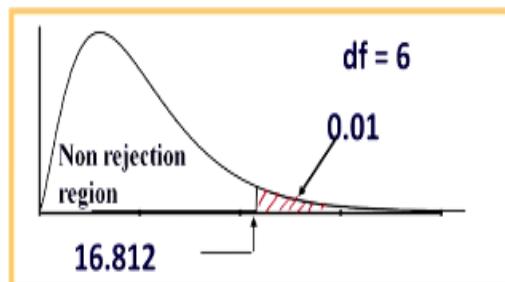
Gasoline Preference Versus Income Category: χ^2 Calculation

$$\begin{aligned}
 \chi^2 &= \sum \sum \frac{(f_o - f_e)^2}{f_e} \\
 &= \frac{(85 - 66.15)^2}{66.15} + \frac{(16 - 24.46)^2}{24.46} + \frac{(6 - 16.40)^2}{16.40} + \\
 &\quad \frac{(102 - 87.78)^2}{87.78} + \frac{(27 - 32.46)^2}{32.46} + \frac{(13 - 21.76)^2}{21.76} + \\
 &\quad \frac{(36 - 45.13)^2}{45.13} + \frac{(22 - 16.69)^2}{16.69} + \frac{(15 - 11.19)^2}{11.19} + \\
 &\quad \frac{(15 - 38.95)^2}{38.95} + \frac{(23 - 14.40)^2}{14.40} + \frac{(25 - 9.65)^2}{9.65} \\
 &= 70.75
 \end{aligned}$$

Now we will find out the chi-square calculated value as we know that the formula for chi-square calculated value is observed frequency – expected frequency divided by expected frequency. So that, you see that, $(85 - 66.15)^2 / 66.15$ it is the first cell first row first cell. First row second cells this 16 is your observed value this is 24.46 is our expected value, whole square divided by 24.46. So if you keep on extent this you can go up to last cell so when you sum it, it is coming 70.75 this is 5 this is 70.75.

(Refer Slide Time: 17:44)

Gasoline Preference Versus Income Category: Conclusion



$$\chi_{Cal}^2 = 70.75 > 16.812, \text{ reject } H_0.$$

We know that previously we have seen that this value is 0.01 our calculated value is 70.75 it is lying on the right hand side so we are going to reject our null hypothesis. When you reject our null hypothesis what was null hypothesis their level of income and the type of fuel they choose are independent. So when you reject it what we are concluding there is a dependency between their level of income and the type of fuel they choose. Generally, it is an assumption

when the level of income is increasing they will go for higher quality of the fuel that was the conclusion.

(Refer Slide Time: 18:28)

Contingency Tables

Contingency Tables

- Useful in situations involving multiple population proportions
- Used to classify sample observations according to two or more characteristics
- Also called a cross-classification table.

The table which we have seen previously it is called a contingency table. It is useful in situations involving multiple population proportions. It is used to classify sample observations according to 2 or more characteristics also called cross-classification table another name for contingency table is cross-classification table.

(Refer Slide Time: 18:52)

Contingency Table Example

Hand Preference vs. Gender

Dominant Hand: Left vs. Right

Gender: Male vs. Female

- 2 categories for each variable, so the table is called a 2×2 table
- Suppose we examine a sample of 300 college students

We will solve one example here. It also is similar to our previous problem there is another example. Suppose we are going to compare the hand preference versus gender. So the dominant hand maybe for some people may be left some people is right. The gender is male versus female we are going to have here hypothesis that is there any connection, is there any

dependency between the gender and their dominant hand. So we have 2 categories for each variable. So this is called 2 cross 2 table. We examine the sample of 300 college students this was the outcome.

(Refer Slide Time: 19:35)

Contingency Table Example

Sample results organized in a contingency table:

sample size = $n = 300$:

120 Females, 12 were left handed

180 Males, 24 were left handed

Hand Preference	Gender		
	Female	Male	
Left	12	24	36
Right	108	156	264
	120	180	300



19

In the rows we have asked are you left hand or right hand dominant hand say left in the column we have captured the gender female and male. There are 300 observations out of 300 observations 120 are female 180 are male. Out of 300, 36 are left handed, 264 are right handed. So the sample result organized in a contingency table. The sample size is n so 120 females, 20 were left handed this one, 180 males, 24 were left handed right.

(Refer Slide Time: 20:20)

Contingency Table Example

$H_0: \pi_1 = \pi_2$ (Proportion of females who are left handed is equal to the proportion of males who are left handed)

$H_1: \pi_1 \neq \pi_2$ (The two proportions are not the same Hand preference is not independent of gender)

- If H_0 is true, then the proportion of left-handed females should be the same as the proportion of left-handed males.
- The two proportions above should be the same as the proportion of left-handed people overall.

So what is our hypothesis $H_0 : \pi_1 = \pi_2$. The proportion of females who are left handed is equal to the proportion of male who are left handed. Now dominant hand left hand is taken as

the reference we are going to compare that left hand people with respect to their gender. So taking left hand is a dominant hand we are going to find out, is there any connection between their hand dominant and their gender that is a null hypothesis.

Null hypothesis proportion of female who are left handed is equal to the proportion of male who are left handed. Suppose if you accept null hypothesis so there is no connection between their dominance of left hand and their gender. What is alternate hypothesis? The two proportions are not the same hand preference is not independent of gender. So what will happen if H_0 is true then the proportion of left handed female should be there.

Same as the proportion of left handed males. So we can say there is no dependency. The two proportions above should be the same as the proportion of left handed people overall instead of taking left hand as a reference you can take the right hand also both result will be the same.

(Refer Slide Time: 21:47)

The Chi-Square Test Statistic

The Chi-square test statistic is:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

where:

f_o = observed frequency in a particular cell

f_e = expected frequency in a particular cell if H_0 is true

χ^2 for the 2×2 case has 1 degree of freedom

Assumed: each cell in the contingency table has expected frequency of at least 5

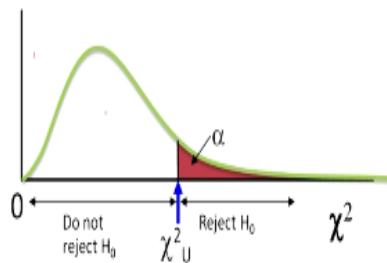
We have seen this formula that is a chi-square test statistics is observed frequency minus expected frequency whole square divided by expected frequency. It says observed frequency f_o is expected frequency. Here there are 2 cross 2 table so the degrees of freedom is $2 - 1$ multiple by $2 - 1$ so it is 1 degrees of freedom. We assume that there is an important assumption each cell in the contingency table has the expected frequencies at least 5. We have to make sure that the expected frequency is at least 5 that is an assumption if it is not there we have to collapse 2, 3 column so that to get the expected frequency is 5.

(Refer Slide Time: 22:36)

The Chi-Square Test Statistic

The χ^2 test statistic approximately follows a chi-square distribution with one degree of freedom

Decision Rule:
If $\chi^2 > \chi^2_U$, reject H_0 ,
otherwise, do not reject H_0



The chi-square test statistic approximately follows the chi-square distribution with one degrees of freedom what will happen the decision rule is if the chi-square value is greater than this limit we will reject it otherwise we will accept it otherwise call it, do not reject it.

(Refer Slide Time: 22:54)

Observed vs. Expected Frequencies

Hand Preference	Gender		
	Female	Male	
Left	Observed = 12 Expected = 14.4 $\frac{36 \times 12}{180}$	Observed = 24 Expected = 21.6 $\frac{36 \times 180}{180}$	36
Right	Observed = 108 Expected = 105.6 $\frac{264 \times 120}{300}$	Observed = 156 Expected = 158.4 $\frac{264 \times 180}{300}$	264
	120	180	300

So this is the observed frequency. Now we have to find out for each cell expected frequency. How we got this expected frequency is nothing but 36 multiplied by 120 divided by 300. So how we got this one 36 multiplied by 124 divided by 300. How we got this value 36 multiple by 180 divided by 300 the same way how we got here 264 multiple by 120 divided not 120 300 here 264 multiplied by 180 divided by 300. So we will get here value for this.

(Refer Slide Time: 23:50)

The Chi-Square Test Statistic

Hand Preference	Gender		
	Female	Male	
Left	Observed = 12 Expected = 14.4	Observed = 24 ✓ Expected = 21.6	36
Right	Observed = 108 Expected = 105.6	Observed = 156 Expected = 158.4	264
	120	180	300

The test statistic is:

$$\chi^2 = \sum_{all cells} \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(12 - 14.4)^2}{14.4} + \frac{(108 - 105.6)^2}{105.6} + \frac{(24 - 21.6)^2}{21.6} + \frac{(156 - 158.4)^2}{158.4} = 0.7576$$

The next one we have to find out the observed frequency minus expected frequency whole square divided by expected frequency plus for this one $(108 - 105.6)^2 / 105.6$ + for this cell it is $(24 - 21.6)^2 / 21.6$, whole square only for the numerator + $(156 - 158.4)^2 / 158.4$ that is we are getting this 0.7576.

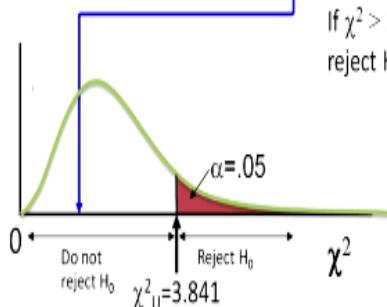
(Refer Slide Time: 24:28)

The Chi-Square Test Statistic

The test statistic is $\chi^2 = 0.7576$, χ^2_U with 1 d.f. = 3.841

Decision Rule:

If $\chi^2 > 3.841$, reject H_0 , otherwise, do not reject H_0



Here,
 $\chi^2 = 0.7576 < \chi^2_U = 3.841$,
so you do not reject H_0 and
conclude that there is
insufficient evidence that the
two proportions are different.

Now we have to mark this one so point the table value which we got this is chi-square calculated value the table values which for one degrees of freedom this is 3.814 our calculated value is lying on the acceptance side. So we have to accept null hypothesis. If chi-square value is greater than 3.841 reject H_0 , otherwise do not reject it here the chi-square value that is 07576 is less than your 3.841.

You do not reject H₀ and conclude that there is insufficient evidence that the 2 proportions are different that means that both are same P₁ = P₂.

(Refer Slide Time: 25:17)

χ^2 Test for The Differences Among More Than Two Proportions

- Extend the χ^2 test to the case with more than two independent populations:

$$H_0: \pi_1 = \pi_2 = \dots = \pi_c$$

$$H_1: \text{Not all of the } \pi_j \text{ are equal } (j = 1, 2, \dots, c)$$

There we have compared only 2 proportions there may be a possibility we have to compare more than 2 proportions for example 3 proportions that case we will see in this problem. Extend the chi-square test to the case where with more than 2 independent populations say null hypothesis can be $\pi_1 = \pi_2 = \pi_3$ the alternate hypothesis not all of the proportions are equal.

(Refer Slide Time: 25:44)

The Chi-Square Test Statistic

The Chi-square test statistic is:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

where:

- f_o = observed frequency in a particular cell of the $2 \times c$ table
- f_e = expected frequency in a particular cell if H_0 is true
- χ^2 for the $2 \times c$ case has $(2-1)(c-1) = c - 1$ degrees of freedom

Assumed: each cell in the contingency table has expected frequency of at least 5

This other formulas are same as usual f_0 is observed frequency f_e is expected frequency chi-square is the degrees of freedom. So the formula number of row – 1 multiplied by number of column – 1.

(Refer Slide Time: 25:58)

χ^2 Test with More Than Two Proportions: Example

The sharing of patient records is a controversial issue in health care. A survey of 500 respondents asked whether they objected to their records being shared by insurance companies, by pharmacies, and by medical researchers.
The results are summarized on the following table:

We will see one example the sharing of patients records is a controversial issue in health care. A survey of 500 respondents asked whether they objected to their record being shared by insurance companies, pharmacies and by medical researchers. The results are summarized on the following table. So there are 3 category now one is whether they are objected to share the data for insurance companies, pharmacies and medical researcher.

(Refer Slide Time: 26:29)

χ^2 Test with More Than Two Proportions: Example

Object to Record Sharing	Organization			Row Sum
	Insurance Companies	Pharmacies	Medical Researchers	
Yes	410 $1040 \times \frac{500}{1040}$	295 $1040 \times \frac{500}{1040}$	335 $1040 \times \frac{500}{1040}$	1040 →
No	90 $1500 - 1040$	205	165	460 →
Column Sum	500	500	500	1500 ↓

So this table shows like this. So you see that 410 patients have objected to share their data with the insurance companies, 295 patients have objected to share their data with the pharmacies, 335 people have objected their data to share with the medical researchers. Now we have to find out whether the proportion of objection for sharing their data all these 3 categories are same or not. So we can say this is π_1 , this is π_2 , this is π_3 .

So null hypothesis will be $\pi_1 = \pi_2 = \pi_3$ that means that the people are always object to share their data irrespective of what kind of company it is that is our null hypothesis. There is a independency between sharing their data and the types of companies which they ask for the data. Here what you have done I have done the row sum this was the row sum this was second row sum then I found the column sum there are 3 columns. This data is our observed frequency 295, 410, 335.

(Refer Slide Time: 27:54)

χ^2 Test with More Than Two Proportions: Example

The overall proportion is:

$$\bar{p} = \frac{X_1 + X_2 + \dots + X_e}{n_1 + n_2 + \dots + n_e} = \frac{410 + 295 + 335}{500 + 500 + 500} = 0.6933$$

Object to Record Sharing	Organization		
	Insurance Companies	Pharmacies	Medical Researchers
Yes	$f_o = 410$ $f_e = 346.667$	$f_o = 295$ $f_e = 346.667$	$f_o = 335$ $f_e = 346.667$
No	$f_o = 90$ $f_e = 153.333$	$f_o = 205$ $f_e = 153.333$	$f_o = 165$ $f_e = 153.333$

Next one from the observed frequency I have to find out the expected frequency. We have already observed frequency how will you find the expected frequency for example here 1040 multiplied by 500 divided by 1,500. So that value it will be 346.667. Similarly, for second dataset it is 1,040 multiplied by 500 divided by 1,500 that data is about 346. This way you can find out the expected frequency.

(Refer Slide Time: 28:40)

χ^2 Test with More Than Two Proportions: Example

Object to Record Sharing	Organization		
	Insurance Companies	Pharmacies	Medical Researchers
Yes	$\frac{(f_o - f_e)^2}{f_e} = 11.571$	$\frac{(f_o - f_e)^2}{f_e} = 7.700$	$\frac{(f_o - f_e)^2}{f_e} = 0.3926$
No	$\frac{(f_o - f_e)^2}{f_e} = 26.159$	$\frac{(f_o - f_e)^2}{f_e} = 17.409$	$\frac{(f_o - f_e)^2}{f_e} = 0.888$

The Chi-square test statistic is:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e} = 64.1196$$

Now I have given the final answer for the first cell that is the observed frequency minus expected frequency whole square divided by expected frequency for this cell it is 11.157, here it is 7.77 this is 0.392, this is 26.159, this is 17.409, this is 0.88. When you add it that value is your 64.1196.

(Refer Slide Time: 29:05)

χ^2 Test with More Than Two Proportions: Example

$$H_0: \pi_1 = \pi_2 = \pi_3 \checkmark$$

$$H_1: \text{Not all of the } \pi_j \text{ are equal } (j = 1, 2, 3)$$

Decision Rule: $\chi^2_U = 5.991$ is from the chi-square

If $\chi^2 > \chi^2_U$, reject H_0 , otherwise, distribution with 2 degrees of freedom. $(2-1)(3-1) = 1 \times 2 = 2$

Conclusion: Since $64.1196 > 5.991$, you reject H_0 and you conclude that at least one proportion of respondents who object to their records being shared is different across the three organizations

So what is a null hypothesis as I told you $\pi_1 = \pi_2 = \pi_3$ alternate hypothesis all of the π_j are equal that is $j : 1, 2, 3, 4$. Decision rules if the calculated chi-square value is greater than your table value reject H_0 otherwise do not reject it. The table value which we got from the table is 5.9 what is the degrees of freedom for knowing this. You see that there are 2 rows is there so $2 - 1$ there are 3 column is there so $3 - 1$.

So this is 1 multiplied by 2 it will be 2 degrees of freedom. For 2 degrees of freedom for given alpha value the chi-square value which you got from the table is 5.991, but you see that our calculated value is our 64.116. So it is bigger than our table value so we have to reject and we can conclude that at least one proportion of the responds to object to their record being shared it is different across the 3 organizations.

So what will happen when you reject to a null hypothesis we can say it is not always equal there are somewhere it is not equal. So not all of the π_j are equal. In this lecture, we started a new topic that is a chi-square test. Chi-square test has 2 applications. One is test of independence and goodness of fit. Today we have started with a test of independence I have taken a small example.

I have explained with the help of example how to test the test of independence. In the next class we will take one small problem with the help of python I will explain how to construct the contingency table. After constructing contingency table how to do chi-square test of independence using python that we will see in the next class. Thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 47
Chi-Square Test of Independence-II

In the last class we started about the chi-square distribution. As I told you chi-square distribution having 2 application one is to test the independence, second one is the goodness of fit. We have seen in the previous class one example of how to test independence of the 2 variables. In this class, we will continue with that we will take another example we will solve with the help of python then we will go for another application of chi-square test that is the goodness of fit.

(Refer Slide Time: 00:56)

Agenda

- Using python to test the independence of variables
- Understanding goodness of fit test for Poisson

So agenda for this lecture is using python to test the independence of the variables. We will start another new topic that is application of chi-square distribution on testing the goodness of fit. First we will test the Poisson distribution.

(Refer Slide Time: 01:13)

Example

- Record of 50 students studying in ABN School is taken at random, the first 10 entries are like this:

res_num	aa	pe	sm	ae	r	g	c
1	99	19	1	2	0	0	1
2	46	12	0	0	0	0	0
3	57	15	1	1	0	0	0
4	94	18	2	2	1	1	1
5	82	13	2	1	1	1	1
6	59	12	0	0	2	0	0
7	61	12	1	2	0	0	0
8	29	9	0	0	1	1	0
9	36	13	1	1	0	0	0
10	91	16	2	2	1	1	0

This is an example record of 50 students studying say ABN school is taken at random. The first 10 entries are shown in like this. Why I have taken this example is in the previous lecture we have got the contingency table directly. The contingency table is given to you once the contingency table is given finding observed frequency and expected frequency then finding chi-square calculated values are very simple.

But in practical the chi-square distribution will not be given to you directly only excel data file or some file from database will be given. You have to form the contingency table after forming the contingency table then you should go for chi-square test. So in this example I have taken one hypothetical problem and there are 1, 2, 3, 4, 5, 6, 7 variables are there. The first variable is academic ability aa.

Second variable is parent education, the third variable is student motivation, the fourth variable is advisory evaluation, the next variable is religion, next variable is gender the last variable is community type. This dataset is to identify what are the variable that affect academic performance of a candidate. So here the academic ability is nothing, but test conducted for out of 100 marks.

So this is marks obtained by the candidates only I have shown only 10 dataset like that there are 50 dataset is there where I am show you when I am showing the python demo. The first one is the marks obtained by a student that is called academic ability so that is higher the marks higher the academic ability. The next variable is parent education the question is asked to the parent how many years you spend on the schooling.

Generally, parent education is a categorical variable, but instead of capturing categorical variable we have got that variable in the form of interval kind of a continuous variable. What we have asked to the parent that how many years you have spent in the schooling say 5 years, 6 years and so on. So now the parent education will become a continuous variable then student motivation.

We have asked the student suppose if you want to study if you want to get more marks are you willing to spend extra time for the study 1 means no, 0 means not decided, 2 means yes advisory evaluation. Advisory is like kind of a faculty advisor that faculty advisor can say whether this fellow will pass in the examination or he will do good performance in the examination.

1 means we will not do 0 means not decided, 2 means we will do good performance in the examination then r is the religion. There are 3 categories of the religion 0, 1, 2 this is the explanation of the variables.

(Refer Slide Time: 04:16)

Example

Here :

- res_num = registration no.
- aa = academic ability
- pe = parent education
- sm = student motivation
- r = religion
- g = gender

For example, as I told you the first column is res respondent underscore number kind of a registration number aa is academic ability, pe is parent education, sm is student motivation, r is the religion, g is the gender.

(Refer Slide Time: 04:35)

Python code

```
In [1]: import pandas as pd  
import numpy as np  
  
In [2]: acad = pd.read_csv('AcademicAbilityData.csv')  
  
In [3]: acad  
Out[3]:
```

	res	sex	aa	pe	sm	ae	r	g	c
0		1	99	19	1	2	0	0	1
1		2	45	12	0	0	0	0	0
2		3	57	15	1	1	0	0	0
3		4	34	19	2	2	1	1	1
4		5	82	13	2	1	1	1	1
5		6	59	12	0	0	2	0	0
6		7	81	12	1	2	0	0	0
7		8	29	9	0	0	1	1	0
8		9	30	13	1	1	0	0	0
9		10	91	95	2	2	1	1	0
10		11	55	93	0	0	1	0	0
11		12	58	11	0	1	0	0	0

I have brought the screenshot of python we have imported pandas and numpy then we imported the dataset so this was the dataset. You can see that there are different variables academic ability, parent education, student motivation, advisory evaluation, religion and gender. So I have imported the dataset the data which have stored in this file name academic ability.data. csv file. So this was the output of the data.

(Refer Slide Time: 05:05)

Hypothesis

- Test the hypothesis that “gender and student motivation” are independent

Now we are going to have the hypothesis. Test the hypothesis that gender and student motivations are independent. Now we are going to see is there any connection between the level of motivation and their gender. In many case we presume that the girls students are highly motivated than the boy students. It is in perception that we will test that, whether is there any connection between any gender and the motivation.

So the null hypothesis is that the gender and student motivations are independent. Alternative hypothesis is it is not independent. The hypothesis which we are going to test is gender and student motivations are independent that is our null hypothesis.

(Refer Slide Time: 05:54)

Python code

```
In [19]: #Cross table between gender and student's motivation
obs = pd.pivot_table(acad[['g','sm']],index = 'g',columns='sm',aggfunc=len)
obs

Out[19]:
sm   0   1   2
g
  0    10  13  6
  1     4   9  8
```

This is very important command for forming our contingency table otherwise call it as a cross table between gender and student motivation. So you see that pd.pivot_table the file name acad, g is the gender, sm student motivation, index g should appear in row column should be the sm, so aggregate function is length. So after you run this we are getting a contingency table. In contingency table here what is there it is a gender. So in column it is a student motivation.

(Refer Slide Time: 06:35)

Observed values

Gender	Student motivation			Row Sum
	0 (Disagree)	1 (Not decided)	2 (Agree)	
0 (Male)	10	13	6	29
1 (Female)	4	9	8	21
Column Sum	14	22	14	50

In the previous table which are shown in here just I have wrote in the presentation. In the row say 0 means it is male 1 means it is a female that is a code which I have used for gender. The student motivation there are 3 level one is 0 disagree. The question which I have asked is are you willing to spend extra time for getting more marks 0 disagree, 1 not decided, 2 agree. So I have seen the row sum there are 29 male students, there are 21 female students.

So 14 students have told that they will not spend extra time, 22 people have decided that told that they have not decided whether they should spend the extra time to study or not. So 14 people have agreed that they will study, they will spend extra time. Now you see that in row there is a one variable in column there is another variable. The null hypothesis is the gender and the student motivations are independent. We know that the cell the 10, 13, 6 represents the observed frequency. Now we have to find out the expected frequency for each cell.

(Refer Slide Time: 07:54)

Expected frequency (contingency table)

Gender	Student motivation		
	0	1	2
0	$29*14/50 =$ 8.12	12.76	8.12
1	5.88	9.24	5.88

The expected frequency is as I told you in the previous class that row total multiplied by column total divided by overall. See previous one 29 multiplied by 14 divided by 50. So this 8.12 is the expected frequency. Similarly, for this also we got 12.76, 8.12, 5.88, 9.24, 5.88. One important assumption that the value of the expected frequency should be more than 5. In all the cells the expected frequency is more than 5 then we can continue for our calculation.

(Refer Slide Time: 08:39)

Frequency Table

Gender	Student motivation		
	0	1	2
0	$f_o = 10$ $f_e = 8.12$	$f_o = 13$ $f_e = 12.76$	$f_o = 6$ $f_e = 8.12$
1	$f_o = 4$ $f_e = 5.88$	$f_o = 9$ $f_e = 9.24$	$f_o = 8$ $f_e = 5.88$

So I have wrote f_o represents observed frequency, f_e represents expected frequency. For each cell we know that what is the observed frequency and expected frequency.

(Refer Slide Time: 08:51)

Chi sq. calculation

$$\chi^2 = \sum \sum \left(\frac{f_o - f_e}{f_e} \right)^2$$

$$= 0.435 + 0.005 + 0.554 + 0.601 + 0.006 + 0.764$$

$$= 2.365$$

Now we have to go for chi-square calculation. We have seen that formula chi-square is calculated value is nothing but observed frequency minus expected frequency whole square divided by expected frequency. So when I go back the first cell observed frequency is $(10 - 8.12)^2$ divided by 8.12 we will get 0.435. Like this when you do for all the cells and sum it the calculated chi-square value is 2.365.

(Refer Slide Time: 09:26)

Python code

```
In [11]: ## Perform chi2 test to check independence
from scipy.stats import chi2_contingency

In [14]: chi2, p,dof,tbl=chi2_contingency(obs)

In [15]: chi2
Out[15]: 2.3649585225939904

In [16]: p
Out[16]: 0.3065178579178871

In [17]: dof
Out[17]: 2
```

Degrees of freedom = $(2-1)*(3-1)$

The python code shows chi-square calculated values for that purpose we have to import chi2_contingency. You see that chi2, p, degrees of freedom, tbl = chi2_contingency the observed values then when you type chi2 you are getting 2.364 the p value is 0.30. Since the p value is more than 0.05 we have accept null hypothesis.

When we say accept null hypothesis we are concluding that the gender and the level of motivations are independent then you can get the degrees of freedom values also. As I told you the degrees of freedom is number of row – 1, 2 – 1 there are 3 columns 3 – 1 say 2 let us say 2 value. This is the degrees of freedom.

(Refer Slide Time: 10:20)

Python code

```
In [12]: M tbl  
Out[12]: array([[ 8.12, 12.76,  8.12],  
                 [ 5.88,  9.24,  5.88]])
```

Contingency table

Then this is contingency table where you can get the expected frequency when you type the tbl you can get the expected frequency you see that this 8.12 this was the I will go back see 8.12, 12.76, 8.12 that value which we got it manually we can directly can get with the help of python. So far I have shown the screenshot of our python programming and explained how to do, how to form a contingency table then how to do the chi-square test. Now I will go to python prompt there I will explain how to input the data and how to do the chi-square test.

(Refer Slide Time: 11:05)

The screenshot shows a Jupyter Notebook interface with several code cells. The first cell imports pandas and numpy. The second cell reads an Excel file named 'acad.xlsx'. The third cell creates a cross-table between gender and student's motivation. The fourth cell performs a Chi-Square test using the scipy.stats module. The fifth cell prints the result, which includes the Chi-Square statistic, degrees of freedom, and p-value.

```
In [1]: import pandas as pd  
import numpy as np  
In [2]: acad = pd.read_excel('acad.xlsx')  
In [3]: acad  
In [4]: # Cross table between gender and student's motivation  
obs = pd.pivot_table(acad[['g','sm']],index = 'g',columns='sm',aggfunc=len)  
obs  
In [5]: # Perform chi2 test to check independence  
from scipy.stats import chi2_contingency  
In [6]: chi2, p, dof,tbl= chi2_contingency(obs)  
In [7]: chi2  
In [8]: p
```

I am going to explain how to form a contingency table from the excel file then doing the chi-square test, import pandas as pd, import numpy as np I have imported, I have imported, I have stored my dataset and the file name called acad.xlsx. So first I have to run this.

(Refer Slide Time: 11:25)

```
In [3]: acad = pd.read_excel('acads.xlsx')
In [4]: acad
Out[4]:
   RipNo aa per sm re g
0      1 99 19 1 2 0 1
1      2 48 12 0 0 0 0
2      3 67 15 1 1 0 0 0
3      4 54 18 2 2 1 1 1
4      5 82 13 2 1 1 1 1
5      6 59 12 0 0 2 0 0
6      7 61 12 1 2 0 0 0
7      8 29 9 0 0 1 1 0
8      9 39 13 1 1 0 0 0
9      10 91 16 2 2 1 1 0
10     11 55 10 0 0 1 0 0
11     12 58 11 0 1 2 0 0
```

Now let us see what is the dataset that dataset you see that there are $49 + 0$ (index) there are 50 dataset is there, that is the respondent number, academic ability, parent education, student motivation, advisory evaluation, religion, gender and community. We are not going to consider all the variables for our calculation. We are just going to consider only the gender and the student motivation okay.

(Refer Slide Time: 11:53)

```
In [5]: #Cross table between gender and student's motivation
obs = pd.pivot_table(acad[['g','sm']], index = 'g',columns='sm',aggfunc=len)
obs
In [6]: # Perform chi2 test to check independence
from scipy.stats import chi2_contingency
In [7]: chi2, p, dof, tb = chi2_contingency(obs)
In [8]: p
In [9]: dof
In [10]: tb
```

Then we will form a contingency table for that. This is $\text{obs} = \text{pd.pivot_table}(\text{academic ability}$ that is the file name, column g and student motivation and index in row I need to have the g value that is the gender value, in column I need to have the student motivation value. So when I see this dataset you see that the output shows that directly I am getting contingency table.

Because when I am explaining theory what happened the contingency table is given to you, but many times that is not the case. The data maybe in some other format you have to create a contingency table before doing the chi-square test. So this command is helping in python, this command is helping us to form the contingency table and it saves lot of our time. So this was the contingency table.

Thus the value in the cell represents the observed frequency. So what this 10 represents when the student level of motivation is 0 the 0 represents male. This is the 10 is our observed value then import chi2_contingency library then we will write chi2, p, degrees of freedom, tbl = chi2_contingency(obs), this obs this obs is wherever contingency table is stored.

So when you run this now the contingency table is run now we want to know the chi-square value the chi-square value is 2.36. This was the chi-square calculated value then we can know the p value, the p value is 0.30 look at the p value which is more than 0.05. So we have to accept our null hypothesis when I say accepting null hypothesis I am concluding that the gender and the student motivations are independent.

There is no connection between the gender and the level of motivation for the student. So we can get to know the degrees of freedom also directly with the help of this command dof and then tb1 this give you your expected frequency. If you are doing manually you can compare this expected frequency. So this was the answer which I have shown in my presentation.

(Refer Slide Time: 14:13)

χ^2 Goodness of Fit Test

So far we have seen the first application of chi-square distribution that is test of independency. Now we are moving into another application that is testing, goodness of fit. What is the meaning of goodness of fit. Many time when we collect the data we have to know what distribution this data follows. So the chi-square test is helping us to find out to know what is the distribution this data follows. First you have to take an example of Poisson data then we will check this whether this data follow Poisson distribution or not.

(Refer Slide Time: 14:48)

χ^2 Goodness-of-Fit Test

- The χ^2 goodness-of-fit test compares *expected* (theoretical) frequencies of categories from a population distribution to the *observed* (actual) frequencies from a distribution to determine whether there is a difference between what was expected and what was observed

What is the chi-square goodness of fit test? Chi-square goodness of fit test compares expected frequencies of categories from population distribution to the observed frequencies, from the distribution to determine whether there is a difference between what was expected and what was observed. So what we are going to do as usual we are also going to see expected frequencies and observed frequencies. We are going to see is there any difference is there or not.

(Refer Slide Time: 15:16)

χ^2 Goodness-of-Fit Test

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$df = k - 1 - p$$

where: f_o = frequency of observed values

f_e = frequency of expected values

k = number of categories

p = number of parameters estimated from the sample data

The formula is same the chi-square value is observed frequency – expected frequency whole square divided by expected frequency. The degrees of freedom this is different from previously what we have seen. In contingency table, the degrees of freedom is number of rows – number of column. Here the degrees of freedom is $k - 1 - p$. The degrees of freedom is $k - 1 - p$ where the k is number of categories number of categories.

Number of observations we can say number of categories, here p is number of parameter estimated from the sample data. This the number of parameters you should know in advance. For example, if it is a uniform distribution parameter is 0 if it is a Poisson distribution the parameter is 1 that is only the lambda value. If it is a normal distribution the parameter is 2 because normal distribution is having 2 parameter one is mean and variance.

(Refer Slide Time: 16:16)

Goodness of Fit Test: Poisson Distribution

1. Set up the null and alternative hypotheses.

H_0 : Population has a Poisson probability distribution

H_a : Population does not have a Poisson distribution

2. Select a random sample and

- Record the observed frequency f_i for each value of the Poisson random variable.
- Compute the mean number of occurrences μ .

3. Compute the expected frequency of occurrences e_i for each value of the Poisson random variable.

Now let us follow some steps to test the goodness of fit of a given dataset. Now assume that some dataset is given to you we are going to test whether this dataset follows Poisson distribution or not. The first step is to setup the null and alternative hypothesis. What is a null hypothesis population has a Poisson probability distribution. What is alternative hypothesis? Population does not have the Poisson distribution.

Here one important point you have to see so far whenever we see the null hypothesis we say that then the term not will appear in the null hypothesis, but only in the goodness of fit test it is reverse. You see that the given data follow Poisson distribution that should be our null hypothesis. Alternative hypothesis is the data the population does not have a Poisson distribution it is just reverse of that.

For all kind of hypothesis or testing the word not will appear in null hypothesis only for the goodness of fit test the word not will appear in your alternative hypothesis. This is one important difference that you have to remember. Select a random sample and record the observed frequency we call as f_i from each value of the Poisson random variable. Compute the mean number of occurrences μ . Because we should know the parameter of Poisson distribution μ compute the expected frequency of occurrences that is e_i for each value of the Poisson random variable.

(Refer Slide Time: 17:58)

Goodness of Fit Test: Poisson Distribution

4. Compute the value of the test statistic

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - e_i)^2}{e_i}$$

where:

f_i = observed frequency for category i

e_i = expected frequency for category i

k = number of categories

Then compute the value of test statistics this is as usual observed frequency – expected frequency whole square divided by expected frequencies. Remember here also the value of expected frequency should be 5 or more. If the expected frequency is not 5 or more you have

to collapse certain intervals and you have to make it so that the expected frequency is 5 or more. We will see that example also here.

(Refer Slide Time: 18:27)

Goodness of Fit Test: Poisson Distribution

5. Rejection rule:

p-value approach: Reject H_0 if p-value $\leq \alpha$

Critical value approach: Reject H_0 if $\chi^2 \geq \chi_{\alpha}^2$

where α is the significance level and
there are k - 2 degrees of freedom

$$k - 1 - p$$

There are 2 way for rejection rule p-value approach. Reject H_0 the p value is $<$ or $=$ alpha if you follow a critical value approach reject H_0 if the chi-square calculated value is greater than your chi-square critical value which you got from the table where the alpha is significance level and there are $k - 2$ degrees of freedom you should remember how this k because it has come $k - 1 - p$. Because Poisson distribution having one parameter only mean is the parameter for the Poisson distribution. So the value of p is 1 so it has become $k - 2$.

(Refer Slide Time: 19:11)

Goodness of Fit Test: Poisson Distribution

- Example: Parking Garage

In studying the need for an additional entrance to a city parking garage, a consultant has recommended an analysis, that approach is applicable only in situations where the number of cars entering during a specified time period follows a Poisson distribution.

So we will take an example see Parking Garage example. In studying need for an additional entrance to a city parking garage, a consultant has recommended an analysis, consultant has

given some solution, that approach is applicable only in situations where the number of cars entering during a specified time period follows Poisson distribution. Since the consultant has given some solution that can be implemented only if the arrival follow Poisson distribution.

(Refer Slide Time: 19:48)

Goodness of Fit Test: Poisson Distribution

A random sample of 100 one- minute time intervals resulted in the customer arrivals listed below. A statistical test must be conducted to see if the assumption of a Poisson distribution is reasonable.

# Arrivals	0	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	0	1	4	10	14	20	12	12	9	8	6	3	1

A random sample of 100 one minute time interval resulted in customer arrival listed below. A statistical test must be conducted to see if the assumption of a Poisson distribution is reasonable. So what is given a number of arrival is given, the frequency is given so 0 arrival the frequency is 0, 1 arrival the frequency is 1, 2 arrival the frequency is 4, 3 arrival frequency is 10 like that up to 12 arrivals is given.

(Refer Slide Time: 20:20)

Goodness of Fit Test: Poisson Distribution

- Hypotheses

H_0 : Number of cars entering the garage during a one-minute interval is Poisson distributed

H_a : Number of cars entering the garage during a one-minute interval is not Poisson distributed

We will form the hypothesis. What is a null hypothesis number of cars entering the garage during one minute interval is Poisson distributed. Alternative hypothesis is number of cars

entering the garage is during a one minute interval is not Poisson distributed. You see that this is different from our traditional way of making null hypothesis. Generally, the term not will be there in the null hypothesis. But here where the goodness of fit test the term the word not will appear in our alternative hypothesis.

(Refer Slide Time: 20:57)

Python Code

```
In [1]: import scipy
from scipy.stats import chisq
from scipy.stats import poisson

In [2]: import pandas as pd
import numpy as np

In [3]: data = pd.read_excel('P_distribution.xlsx')
data
```

Arrivals	Frequency
0	0
1	1
2	4
3	10
4	16
5	20
6	32
7	32
8	9
9	6
10	6
11	3
12	1

This is the python code which I have taken the screenshot import scipy, import chi-square, import Poisson. The dataset I have kept in the file name called P_distribution. This was the arrival this is the actual frequency otherwise we can call it as observed frequency.

(Refer Slide Time: 21:21)

Goodness of Fit Test: Poisson Distribution

- Estimate of Poisson Probability Function

$$\text{Total Arrivals} = 0(0) + 1(1) + 2(4) + \dots + 12(1) = 600$$

$$\text{Estimate of } \mu = 600/100 = 6$$

$$\text{Total Time Periods} = 100$$

Hence,

$$= \frac{e^{-\mu} \mu^x}{x!}$$

$$f(x) = \frac{6^x e^{-6}}{x!}$$

$$\begin{aligned} \mu &= \sum f_n = \frac{600}{100} \\ &= 6 \end{aligned}$$

The next term we should know mean of the dataset. Estimate of Poisson probability function see the total arrival the mean formula we know that the same simple formula mean is mean mu = sigma fn / sigma f. So f is the frequency n is the number of arrival so 0 into 0 + 1 into 1

like that there will be a value will be 600. So sigma f when I go back you see that this frequency when you add this frequency when you add this frequency that will be 100.

So that value is 6 so the 6 is your mean. We know that our formula traditionally what is our formula. So the formula is $(e^{-\mu} \mu^x)/x!$, otherwise some people call it lambda, $(e^{-\mu} \mu^x)/x!$. So mu is 6 so 6 to the power x e to the power - 6 / x factorial.

(Refer Slide Time: 22:44)

Goodness of Fit Test: Poisson Distribution

- Expected Frequencies

x	f(x)	nf(x)	x	f(x)	nf(x)
0	.0025	.25	7	.1377	13.77
1	.0149	1.49	8	.1033	10.33
2	.0446	4.46	9	.0688	6.88
3	.0892	8.92	10	.0413	4.13
4	.1339	13.39	11	.0225	2.25
5	.1606	16.06	12+	.0201	2.01
6	.1606	16.06	Total	1.0000	100.00

Now what we have to do we have to substitute these x values then if you substitute this x values you will get a theoretical frequency. So when x = 0, f(x) is when you substitute in this equation when x = 0 6 to the power 0 e to the power - 6 / 0 factorial that is 1. So e to the power - 6 is this 0.0025 this is probability value. We want to know in terms of frequency so that has to be multiple by n, n is 100.

When you substitute x = 1 in this equation 6 to the power 1 e to the power - lambda / 1 factorial will get 0.0149 then multiple by n. So this value will give you the theoretical frequency of Poisson distribution like that we have to have up to 12. You see that here, here the theoretical frequency when you look at the 0, 1, 2 this values. See that this is less than 5 so this has to be added this also less than 5 so these 3 groups has to be grouped.

The same way you see that this is 6.8 here what is happening these values are less than 5. So these values has to be clubbed so that the expected frequency is 5 that is what we have done that one.

(Refer Slide Time: 24:15)

Python code

```
[In [4]]: (observed_freq * data['Frequency'])

[In [5]]: total_arrival = 600
          total_time_period = 100
          mu = total_arrival/total_time_period

[In [6]]: expected_freq = []
for i in range(len(observed_freq)):
    E_freq = 100*pmisson.pdf(i, mu)
    Expected_freq.append(E_freq)

[In [7]]: Expected_freq
Out[7]: [0.247352176667584,
 1.48751186998045,
 4.46175391799446,
 8.923507815938166,
 13.385361751994372,
 16.062314184197995,
 16.06231418419801,
 13.767079700112569,
 10.3257715188447,
 6.883849920562646,
 4.130109311213764,
 2.3528466031699247,
 1.164488021546481]
```

Okay so we have got the observed frequency we got that mean value then for each x value we got our expected frequency value by using the for loop okay.

(Refer Slide Time: 24:28)

Python code

```
[In [4]]: (expected_freq, round off) = [(round(x, 2)) for x in (expected_freq)]
(expected_freq, round off)

Out[4]: [[6.21,
  3.81,
  1.49,
  0.81,
  0.31,
  0.11,
  0.06,
  0.03,
  0.01,
  0.01],
 [0.247352176667584,
  1.48751186998045,
  4.46175391799446,
  8.923507815938166,
  13.385361751994372,
  16.062314184197995,
  16.06231418419801,
  13.767079700112569,
  10.3257715188447,
  6.883849920562646,
  4.130109311213764,
  2.3528466031699247,
  1.164488021546481]]
```

```
[In [5]]: df = pd.DataFrame([[i,(observed_freq, expected_freq,round off)), i+1] for i in range(len(observed_freq), len(expected_freq))], columns = ['Interval Frequency', 'Expected Frequency'])
df
```

	Interval Frequency	Expected Frequency
0	0	4.21
1	1	3.81
2	1	1.49
3	0	0.81
4	14	0.31
5	19	0.11
6	12	0.06
7	12	0.03
8	9	0.01
9	1	0.01
10	6	0.01
11	3	0.01
12	1	0.01

So then we will do round off this was our rounded value using python. When you look at this 0.25 you go back 0.25, 1.49 you will get exact the same value. Now we are going to have only 2 column one is observed frequency next one is expected frequency.

(Refer Slide Time: 24:48)

Goodness of Fit Test: Poisson Distribution

- Observed and Expected Frequencies

i	f_i	e_i	$f_i - e_i$
0 or 1 or 2	5	6.20	-1.20
3	10	8.92	1.08
4	14	13.39	0.61
5	20	16.06	3.94
6	12	16.06	-4.06
7	12	13.77	-1.77
8	9	10.33	-1.33
9	8	6.88	1.12
10 or more	10	8.39	1.61

You see that 0 or 1 or 2 so that are clubbed so that the expected frequency is more than 5. Similarly, 10, 11, 12 these are grouped together so that the expected frequency is 8.39 otherwise it will be less than 5. Now how many numbers of interval is there 1, 2, 3, 4, 5, 6, 7, 8, 9 interval is there.

(Refer Slide Time: 25:15)

Python code

```
In [10]: obs_freq = [5, 10, 14, 20, 12, 12, 9, 8, 10]
expected_freq = [6.20, 8.92, 13.39, 16.06, 16.06, 13.77, 10.33, 6.88, 8.39]
```

```
In [11]: scipy.stats.chisquare(obs_freq, expected_freq)
```

```
Out[11]: Power_divergenceResult(statistic=3.2738182931105193, pvalue=0.916017731732134)
```

Now this is observed frequency expected frequency. Now if you directly you can run this command that is `scipy.stats.chisquare()`, observed frequency – expected frequency that we are getting 3.27 the p value is 0.911 that is more than 0.05. So we have to accept null hypothesis when we accept null hypothesis we are concluding that the given arrival pattern follow Poisson distribution.

(Refer Slide Time: 25:45)

Goodness of Fit Test: Poisson Distribution

- Rejection Rule

With $\alpha = .05$ and $k - p - 1 = 9 - 1 - 1 = 7$ d.f.

(where k = number of categories and p = number of population parameters estimated),

$$\chi^2_{0.05} = 14.067$$

Reject H_0 if $p\text{-value} \leq .05$ or $\chi^2 \geq 14.067$.

- Test Statistic



$$\chi^2 = \frac{(-1.20)^2}{6.20} + \frac{(1.08)^2}{8.92} + \dots + \frac{(1.61)^2}{8.39} = 3.268$$

You see that for rejection rule when alpha = 0.05, k we have got k = 9 as I told you because 9 interval and going back I will explain 1, 2, 3, 4, 5, 6, 7, 8, 9 interval that is why here k = 9. P is the number of parameter as we know that that Poisson distribution having only one parameter so it is $9 - 1 - 1$ so 7 degrees of freedom. For 7 degrees of freedom when alpha = 0.05 we can get the chi-square table value is 14.06.

See when you look at the chi-square calculated values it is 3.268. So this value will lie on the acceptance side. So we have to accept because the chi-square distribution to be like this so this is 14.067 you are 3.268 will be here it will be lying on the acceptance side so you have to accept the null hypothesis.

(Refer Slide Time: 26:44)

Python code

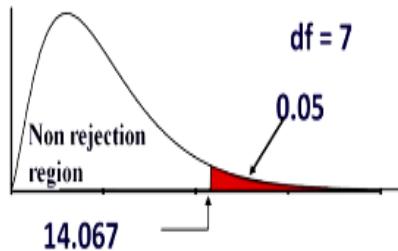
```
In [4]: from scipy.stats import chi2  
chi2.ppf(0.95, 7)
```

```
Out[4]: 14.067140449340167
```

The same thing 0.95, 7 so the chi-square calculated value is 14.06.

(Refer Slide Time: 26:52)

Goodness of Fit Test: Poisson Distribution



$$\chi^2_{cal} = \underline{3.268} < 14.067, \text{ do not reject } H_0.$$

See there 14.06 but not calculated value. The chi-square table value is 14.06, but our calculated value is 3.268 so it is lying on the acceptance side do not reject null hypothesis then we will conclude that the arrival pattern follow Poisson distribution. In this class, I have explained how to form a contingency table after forming contingency table how to do the chi-square test with the help of python.

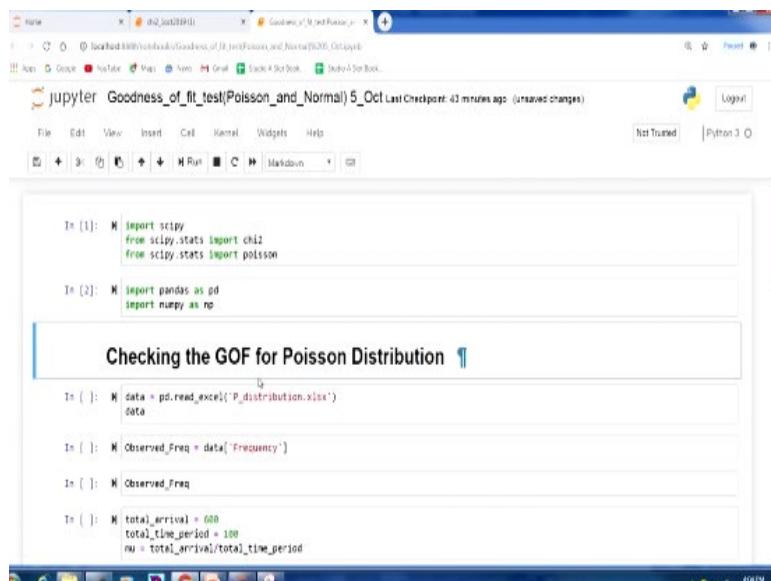
The next topic which I have started testing goodness of fit. Suppose some dataset is given if you want to test what distribution it follows. I have taken some dataset then I have tested whether this dataset follow Poisson distribution or not. I have explained the python screenshot. In the next class I will run the python code for testing Poisson distribution for given dataset and I will explain how to test goodness of fit for uniform distribution and normal distribution that we will see in the next class. Thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 48
Chi Square Goodness of Fit Test

Welcome students. In the previous lecture I have explained how to do goodness of fit for your Poisson distribution. I have explained the theory portions. In this class I am going to give you a python demo for that, after the python demo I am going to explain how to do the goodness of fit for uniform distribution or normal distribution.

(Refer Slide Time: 00:53)



```
In [1]: import scipy
from scipy.stats import chi2
from scipy.stats import poisson

In [2]: import pandas as pd
import numpy as np

Checking the GOF for Poisson Distribution

In [3]: data = pd.read_excel('P_distribution.xlsx')
data

In [4]: Observed_Freq = data['Frequency']

In [5]: Observed_Freq

In [6]: total_arrival = 600
total_time_period = 100
mu = total_arrival/total_time_period
```

The agenda for this lecture is python demo for testing goodness of it for the Poisson distribution, theory of how to do the goodness of fit for uniform or normal distribution. After that I will give you the python demo for testing goodness of fit for uniform and normal distribution. Now we will go to the python prompt there we will see how to do goodness of fit for your Poisson distribution.

Now we will see the python demo say some dataset is given we are going to test whether that dataset follow Poisson distribution or not. So I am importing the necessary library like scipy from scipy.stats chi2, from scipy.stats, poisson then for I am importing pandas and numpy.

(Refer Slide Time: 01:43)

```
In [3]: data = pd.read_excel('P_distribution.xlsx')
data
```

Arrival	Frequency
0	0
1	1
2	4
3	10
4	14
5	20
6	12
7	12
8	9
9	8
10	6
11	3
12	1

So checking goodness of fit for your Poisson distribution we will see what is that dataset. This dataset is given arrival is given, the frequency is given. The arrival is in one minute arrival. For example, in one minute interval 0 interval there is 0 frequency. In one minute interval there is 1 arrival there is 1 frequency. In one minute interval there are 2 interval there are 4 frequency and so on.

(Refer Slide Time: 02:11)

```
In [4]: Observed_Freq = data['Frequency']
```

```
In [5]: Observed_Freq
```

	Frequency
0	0
1	1
2	4
3	10
4	14
5	20
6	12
7	12
8	9
9	8
10	6
11	3
12	1

```
In [6]: total_arrival = 600
total_time_period = 100
mu = total_arrival/total_time_period
```

Whatever the data which is given that is observed frequency. So the observed frequency I am going to call it separately these are our observed frequency. Then next one is we have to find out the expected frequency for given x values. To know the expected frequency, we have to know the mean of the Poisson distribution. We know that mean of the Poisson distribution like $\sigma f n$ divided by σf . See the total arrival is 600 total time period is 100 so 600 divided by 100 so the mu value is your 6.

(Refer Slide Time: 02:58)

```
In [6]: mu = 6  
total_arrival = 600  
total_time_period = 100  
mu = total_arrival/total_time_period  
  
In [7]: mu  
Out[7]: 6.0  
  
In [8]: Expected_Freq = []  
for i in range(len(Observe_Freq)):  
    E_Freq = 100*poisson.pmf(i, mu)  
    Expected_Freq.append(E_Freq)  
  
In [9]: Expected_Freq  
  
In [10]: Expected_Freq_round_off = [round(elem, 2) for elem in Expected_Freq]  
Expected_Freq_round_off  
  
In [11]: df = pd.DataFrame(list(zip(Observe_Freq, Expected_Freq_round_off)), columns = ['Observed Frequency', 'Expected Frequency'])
```

So we know the mu value, we know the x value now we have to find out the expected frequency. So I am making expected underscore frequency so for i in range of length observed frequency then finding the expected frequency. So $E_{frequency} = 100 \times \text{Poisson PMF}(i, \mu)$ then I am going to get the expected frequency. So I am using a for loop so that it will save our time.

(Refer Slide Time: 03:41)

```
In [9]: Expected_Freq  
Out[9]: [0.347521786663584,  
1.48751059981245,  
4.461753012999446,  
8.93567835908894,  
13.30261753928132,  
16.05234164797095,  
16.05234164797095,  
13.78797804122569,  
10.357735300442,  
6.8134890205284,  
4.13930934123764,  
2.752860043095247,  
1.2364840021546681]  
  
In [10]: Expected_Freq_round_off = [round(elem, 2) for elem in Expected_Freq]  
Expected_Freq_round_off  
  
In [11]: df = pd.DataFrame(list(zip(Observe_Freq, Expected_Freq_round_off)), columns = ['Observed Frequency', 'Expected Frequency'])  
df  
  
In [12]: obs_freq = [5, 10, 14, 20, 12, 11, 9, 8, 10]  
expected_freq = [6.0, 8.92, 13.19, 16.00, 16.00, 13.77, 10.51, 6.88, 8.39]
```

So this was our expected frequency. This expected frequency there are different decimal is there suppose I want to round it off to 2 decimal for that purpose you have to use this command equal to square bracket round element, 2 for element in expected frequency. So expected frequency rounded off let us see what is that value.

(Refer Slide Time: 04:05)

```

In [10]: Expected_Freq_round_off = [round(elem, 2) for elem in Expected_Freq]
Expected_Freq_round_off
Out[10]: [0.25,
 1.49,
 4.46,
 8.92,
 13.19,
 16.06,
 16.06,
 13.77,
 10.33,
 6.08,
 4.13,
 2.25,
 1.13]

In [11]: df = pd.DataFrame(list(zip(Observed_Freq, Expected_Freq_round_off)), columns = ['Observed Frequency', 'Expected Frequency'])
df
Out[11]: Observed Frequency  Expected Frequency
0                      0             0.25
1                      1             1.49
2                      4             4.46
3                     10             8.92
4                     14            13.19
5                     20            16.06
6                     12            16.06
7                     12            13.77
8                      9            10.33
9                      8             6.08

```

So this is our expected frequency rounded value. Now we will add these both the variables by using zip command into the object called df.

(Refer Slide Time: 04:19)

```

In [11]: df = pd.DataFrame(list(zip(Observed_Freq, Expected_Freq_round_off)), columns = ['Observed Frequency', 'Expected Frequency'])
df
Out[11]: Observed Frequency  Expected Frequency
0                      0             0.25
1                      1             1.49
2                      4             4.46
3                     10             8.92
4                     14            13.19
5                     20            16.06
6                     12            16.06
7                     12            13.77
8                      9            10.33
9                      8             6.08

```

So the df says that our observed frequency and expected frequency. Once we know the observed frequency and expected frequency then we can get the chi square value directly, but look at here, here the expected frequency is 0.25 that is less than 5. So, up to 3 when you add this 3 then only your expected frequency will be more than 5. So we have added this so after adding so the observed frequency is 4,1,5 the expected frequency is 6.20.

This we have done it manually when you are running a large program with huge dataset you can make the program so that the expected frequency is more than 5, but here we have not

done that way. We have manually added and checked whether the expected frequency is more than 5 or not.

(Refer Slide Time: 05:13)

The screenshot shows a Jupyter Notebook interface with the title "jupyter Goodness_of_fit(Poisson_and_Normal) 5_Oct Last Checkpoint at hour ago (autosaved)". The notebook contains the following code and output:

```
In [13]: obs_freq = [5, 10, 14, 10, 12, 9, 8, 10]
expected_freq = [6.06, 8.92, 13.39, 16.06, 16.06, 13.77, 10.33, 6.88, 8.39]

In [14]: scipy.stats.chisquare(obs_freq, expected_freq)
Out[14]: Power_divergenceResult(statistic=3.273818293105109, pvalue=0.914017731732334)
```

A blue box highlights the text "Checking the GOF for Uniform Distribution".

So this is observed frequency then expected frequency. Now similarly you see that for 10 and 11 and 12 the expected frequency is not more than 5 so we have collapsed these 3 intervals and made into one interval or that is called 10 or more so that is your 8.39 and the interval is 10 or 11 or 12. Now we have the observed frequency and expected frequency. Simply you pass this command that is `scipy.stats.chisquare` observed frequency and expected frequency.

I have to run this. Now I am getting the test statistics chi square statistics 3.27 see the p value is 0.91 it is above our 0.045. So we have to accept our null hypothesis in our presentation also I have told you the p value is 0.91 the calculated chi square value is 3.27. So we have to accept our null hypothesis and we have to conclude that the given dataset follows Poisson distribution.

(Refer Slide Time: 06:22)

Goodness of fit for Uniform Distribution

Milk Sales Data	Month	Litres
	January	1,610
	February	1,585
	March	1,649
	April	1,590
	May	1,540
	June	1,397
	July	1,410
	August	1,350
	September	1,495
	October	1,564
	November	1,602
	December	1,655
		18,447

We have tested whether the given dataset follow Poisson distribution or not. So Poisson distribution is the discrete distribution. We will see another example where the given dataset follow uniform distribution or not. So the milk sales is given for 12 months January, February, March, April, May, June up to Decembers then liters also given like this So we are going to say that the sales follow uniform distribution or not. So the assumption is the sales follow uniform distribution.

(Refer Slide Time: 06:56)

Hypotheses and Decision Rules

H_0 : The monthly milk figures for milk sales are uniformly distributed

H_a : The monthly milk figures for milk sales are not uniformly distributed

$$\begin{aligned}
 \alpha &= .01 && \text{If } \chi^2_{\text{cal}} > 24.725, \text{ reject } H_0. \\
 df &= k - 1 - p && \text{If } \chi^2_{\text{cal}} \leq 24.725, \text{ do not reject } H_0. \\
 &= 12 - 1 - 0 && \\
 &= 11 && \\
 \chi^2_{.01,11} &= 24.725 &&
 \end{aligned}$$

So what is a null hypothesis the monthly milk figures for milk sales are uniformly distributed. Alternative hypothesis is the monthly milk figures of milk sales are not uniformly distributed. You see that the not appears in our alternative hypothesis which is not our traditional way of forming the null hypothesis. So we have alpha = 0.01 the k is the given dataset because there are 12 month dataset is there 12.

And here see the p value is 0 because the uniform distribution is not having any parameter because if you know the lower limit and upper limit of uniform distribution you can easily construct the uniform distribution. So for uniform distribution having the 0 parameter after simplifying it is 11 so when alpha = 0.01 the degrees of freedom is 11, the calculated chi square value is 24.725.

So if we are using the critical value method if the calculated value is greater than not calculated value I am correcting this is the table value, table value which we got from the chi square table 24.725. If the calculated chi square value is greater than 24.725 we have to reject it otherwise we have to accept our null hypothesis.

(Refer Slide Time: 08:24)

Python code

```
In [1]: from scipy.stats import chi2
In [2]: import pandas as pd
         import numpy as np
In [3]: chi2.ppf(0.99,11)
Out[3]: 24.724970311318277
```

We can find out the chi square table value by using this one $\text{chi2.ppf}(0.99, 11)$, 24.72 see that the same value 24.72.

(Refer Slide Time: 08:38)

Calculations

Month	f_o	f_e	$(f_o - f_e)^2/f_e$	
January	1,610	1,537.25	3.44	
February	1,585	1,537.25	1.48	
March	1,649	1,537.25	8.12	
April	1,590	1,537.25	1.81	
May	1,540	1,537.25	0.00	$f_e = \frac{18447}{12} = 1537.25$
June	1,397	1,537.25	12.80	
July	1,410	1,537.25	10.53	
August	1,350	1,537.25	22.81	$\chi^2_{Cal} = 74.37$
September	1,495	1,537.25	1.16	
October	1,564	1,537.25	0.47	
November	1,602	1,537.25	2.73	
December	1,655	1,537.25	9.02	
	18,447	18,447.00	74.38	

Now this is the given dataset month is given observed frequency is given. To know the expected frequency what you have to do we have to sum this dataset see that this is the sum value 18,447 then because it is followed uniform distribution you divide by 12. So, equal value is 1,537.25. So everywhere you can write this should be our expected frequency. Then you find out observed frequency minus expected frequency whole square divided by expected frequency. Then you sum it you are getting 74.38. So the calculated chi square value is 74.38.

(Refer Slide Time: 09:32)

Python code

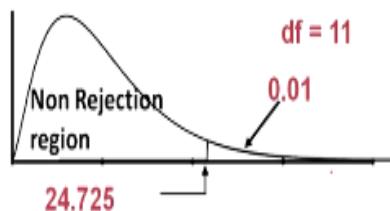
```
In [6]: x =[1610,1585,1649,1590,1540,1397,1410,1350,1495,1564,1602,1655]
In [7]: np.mean(x)
Out[7]: 1537.25
In [8]: exp_f=[1537.25,1537.25,1537.25,1537.25,1537.25,1537.25,1537.25,1537.25,1537.25,1537.25,1537.25,1537.25]
In [9]: from scipy.stats import chisquare
chisquare(x,exp_f)
Out[9]: Power_divergenceResult(statistic=74.37503346885673, pvalue=1.78545252783034e-11)
```

Let us see the python code for this the x is given we are finding the mean 1,537. So expected frequency is nothing but the same value expected frequency this also we have entered manually then from `scipy.stats import chisquare`. So chi square x, expected frequency so, when you give chi square x, expected underscore frequency. You will get the calculated chi

square value is 74.37 the p value is 1.7 into 10 to the power – 11. So you see that this our python outputs and our calculated values are same.

(Refer Slide Time: 10:19)

Conclusion



$$\chi^2_{cal} = 74.37 > 24.725, \text{ reject } H_0.$$

So obviously this was the table value 24.725 our calculated value is this one. So we have to reject our null hypothesis. When we reject a null hypothesis what we are concluding that the dataset is not following uniform distribution.

(Refer Slide Time: 10:41)

Goodness of Fit Test: Normal Distribution

1. Set up the null and alternative hypotheses.
2. Select a random sample and
 - a. Compute the mean and standard deviation.
 - b. Define intervals of values so that the expected frequency is at least 5 for each interval.
 - c. For each interval record the observed frequencies
3. Compute the expected frequency, e_j , for each interval.

Now we will go to another very interesting example testing some dataset and checking whether it is following normal distribution or not. So what are the different steps are there. The first step is setup null and alternative hypothesis, select the random sample and compute the mean and standard deviation. Define intervals of values so that the expected frequency is

at least 5 for each interval. For each interval record the observed frequency then compute the expected frequency for each interval.

(Refer Slide Time: 11:19)

Goodness of Fit Test: Normal Distribution

4. Compute the value of the test statistic.

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - e_i)^2}{e_i}$$

5. Reject H_0 if $\chi^2 \geq \chi_{\alpha}^2$ K-1-P
(where α is the significance level and there are k - 3 degrees of freedom)

Then compute the value of test statistics then reject H_0 if the chi square value is greater than your value which you got from the table. Here the alpha is significance level here the degrees of freedom is $k - 3$ because we know that this is $K - 1 - P$. So here P is 2 because there are 2 parameter for a normal distribution so the value of $P = 2$ that is why we have got $k - 3$.

(Refer Slide Time: 11:53)

Normal Distribution Goodness of Fit Test

- Example: IQL Computers

IQL Computers manufactures and sells a general purpose microcomputer. As part of a study to evaluate sales personnel, management wants to determine, at $\alpha = 0.05$ significance level, if the annual sales volume (number of units sold by a salesperson) follows a normal probability distribution.

We will take an example see computer manufacturers and sells a general purpose microcomputer. As part of a study to evaluate the sales personnel management wants to determine that alpha = 5% significance level if the annual sales volume that is the number of units sold by the sales person follows normal probability distribution. So there are some

dataset that dataset is nothing, but the number of units sold by the sales people. They want to test whether that sales follow normal distribution or not.

(Refer Slide Time: 12:30)

Normal Distribution Goodness of Fit Test

A simple random sample of 30 of the salespeople was taken and their numbers of units sold are below.

```
33 43 44 45 52 52 56 58 63 64  
64 65 66 68 70 72 73 73 74 75  
83 84 85 86 91 92 94 98 102 105
```

(mean = 71, standard deviation = 18.23)

A simple random sample of 30 of the salespeople who has taken and their number of units sold are given below. 33, 43, 44 and so on. So for this dataset the mean is 71, standard deviation is 18.23.

(Refer Slide Time: 12:49)

Python code

```
In [11]: A = [33, 43, 44, 45, 52, 52, 56, 58, 63, 64, 64, 65, 66, 68, 70, 72, 73, 73, 74, 75, 83, 84, 85, 86, 91, 92, 94, 98, 102, 105]  
In [13]: mean = np.mean(A)  
mean  
Out[13]: 71.0  
In [14]: std = np.std(A)  
std  
Out[14]: 18.226154544998845
```

So we have imported the data so we are finding the mean is 71, standard deviation is 18.22.

(Refer Slide Time: 12:57)

Normal Distribution Goodness of Fit Test

- Hypotheses

H_0 : The population of number of units sold
has a normal distribution with mean 71
and standard deviation 18.23

H_a : The population of number of units sold
does not have a normal distribution with
mean 71 and standard deviation 18.23

For this dataset what is the null hypothesis. The population of number of units sold has a normal distribution with mean 71 and standard deviation is 18.23. The alternative hypothesis is the population of number of unit sold does not have a normal distribution with mean 71 and standard deviation 18.23 because many times whenever you collect the data we have to test what distribution it follows.

Because when you do the simulation rather purpose we have to knowing the exact distribution of given dataset is more important that is why testing the particular distribution will be very helpful for further analysis of our dataset.

(Refer Slide Time: 13:47)

Normal Distribution Goodness of Fit Test

- Interval Definition

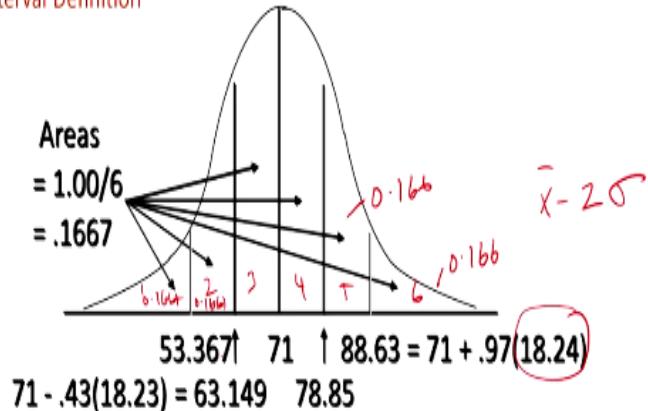
To satisfy the requirement of an expected frequency of at least 5 in each interval we will divide the normal distribution into $30/5 = 6$ equal probability intervals.

First we make an interval to satisfy the requirement of an expected frequency of at least 5 this is very important at least 5 in each interval we will divide the normal distribution by 5 so 30 divided by 5 so that we will get 6 equal probability interval.

(Refer Slide Time: 14:06)

Normal Distribution Goodness of Fit Test

- Interval Definition



You see that 1, 2, 3, 4, 5, 6. So when you divide by 5 we will get 6 interval because then when you divide this way then we can make sure that in every interval you will get minimum 5 or more observed value. So total area = 1 so when you divide 1 divided by 6 we will get 0.1667 is because this area is 0.1667 this area is 0.1667 everywhere. So when the area is 0.167 you can get the z value. We know that their lower limit $X\bar{ } - Z\sigma$.

So $X\bar{ }$ is given Z value you can get it sigma value also given you can find out this interval so this is 53.367 the second point is we know $X\bar{ }$ is 71 how we got the 71 because the mean is 71 right. The mean is 71 the sigma is given, sigma is 18.23 when it is $0.66 + 0.66$ the corresponding Z value is 0.43 so 0.43 and this sigma value that is your 16.17. Similarly, at the right hand side see that how we got 88.63.

The mean value is given so when this side area is 0.166 this area is 0.166 when you add that then corresponding Z value is 0.97. So $71 + 0.97$ this was our sample standard deviation the sigma value then you will get 88.63. What we got it we got different intervals.

(Refer Slide Time: 16:01)

Python code

```
In [15]: x = 1/6 #for 6 equal probability intervals.
```

```
In [16]: for j in range(1,6):
    Prob_intervals = [scipy.stats.norm.ppf(j*x, mean, std)]
    print (Prob_intervals)
```

```
[53.36743154175236]
[63.14941153083116]
[71.0]
[78.85058846916884]
[88.63256845824763]
```

So what are those intervals you see that 1 divided by 6 so that we will get different 6 equals probability intervals. So for j in range 1, 6 the probability underscore interval is `scipy.stats.norm.ppf` we can substitute j value directly here then x , mean in standard deviation. So when you print this probability interval you are getting this is your different intervals when we got after solving manually you see 53.67, the second value is 63.149 this was our interval of the normal distribution.

(Refer Slide Time: 16:43)

Normal Distribution Goodness of Fit Test

- Observed and Expected Frequencies

i	f_i	e_i	$f_i - e_i$
Less than 53.02	6	5	1
53.02 to 63.03	3	5	-2
63.03 to 71.00	6	5	1
71.00 to 78.97	5	5	0
78.97 to 88.98	4	5	-1
More than 88.98	6	5	1
Total	30	30	

$$\frac{30}{6} = 5$$

See that the first value is 0 to 53.36 see that was this value the second one is 53.02, 63.03, 63.14 then 71, 78, 88 the maximum value is 88.63. So what we have to do now we have to go to this values in that interval we have to count it how many numbers are appearing. For example, in the interval you see that when the interval less than 53.02 this 6 was observed frequency.

So how we got the 6 one from this given dataset you have to find out how many numbers are below that when you count it, it will be 6. Similarly, in the interval 53.02, 63.03 in our given dataset we have to count it how many numbers are appearing this range there will be a 3 this is our observed frequency. This is 5 is expected frequency because there was a 30 dataset since we divided 30 divided by 6 there will be a 5 expected frequency will be there.

Because why we are dividing by 5 because minimum we need to have 5 expected frequency that is why we have divided by 6 so we have to divide the given dataset by a number so that the expected frequency is 5. So we got this observed then expected frequency then find the different square the difference.

(Refer Slide Time: 18:26)

Python code

```
In [17]: Expected_Freq = [5,5,5,5,5] #will divide the normal distribution into 6 intervals at frequency 5 in each
```

```
In [18]: Obs_f = [6,3,6,5,4,6]
```

```
In [19]: scipy.stats.chisquare(Obs_f, Expected_Freq)
```

```
Out[19]: Power_divergenceResult(statistic=1.599999999999999, pvalue=0.9012493445012737)
```

$$\alpha = 5\%$$

This was our expected frequency this is our observed frequency. So `scipy.stats.chisquare` observed frequency, expected frequency. We are getting 1.5 see the p value is 0.90 that is we say that 5% it is more than that. We say we have to accept our null hypothesis so the given dataset follow normal distribution.

(Refer Slide Time: 18:50)

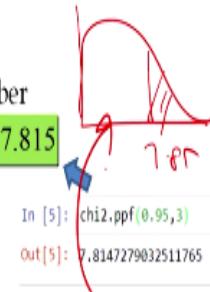
Normal Distribution Goodness of Fit Test

- **Rejection Rule**

With $\alpha = .05$ and $k - p - 1 = 6 - 2 - 1 = 3$ d.f.

(where k = number of categories and p = number of population parameters estimated), $\chi^2_{0.05} = 7.815$

Reject H_0 if $p\text{-value} \leq .05$ or $\chi^2 \geq 7.815$.



- **Test Statistic**

$$\chi^2 = \frac{(1)^2}{5} + \frac{(-2)^2}{5} + \frac{(1)^2}{5} + \frac{(0)^2}{5} + \frac{(-1)^2}{5} + \frac{(1)^2}{5} = 1.600$$

When you look at that one alpha = 0.05 there are k = 6. How we get got is 6 number of intervals I am going back 1, 2, 3, 4, 5, 6 that is why it is k = 6. P is a 2 parameter n so $6 - 2 - 1$ is the 3 degrees of freedom. When 3 degrees of freedom when alpha = 5% the table value is 7.8815 you look at this one our test statistic observed frequency minus expected frequency whole square divided by expected frequency when we add it is 1.6.

So what is happening in chi square distribution so this value is 7.815 our calculated value we are getting 1.6 so we have to accept our null hypothesis and we have to conclude that the given dataset follow normal distribution. Now I will explain the python code of testing how to check whether the given dataset follow uniform distribution and normal distribution.

(Refer Slide Time: 20:01)

Suppose we want to know the chi square table value when alpha equal to 1 percentage when it is alpha 1 percentage we have to write chisquare.ppf (0.99, 11) degrees of freedom. So we are getting the calculate the table value of chi square value is 24.72. Then next one this is our x value then we are finding the mean of that one. So the expected frequency is nothing but it is going to be our mean.

Now we got the observed frequency that is our 1, 610 that is x value than expected frequency is this one exp underscore f. So when you write chi square x, expected underscore f this is our expected frequency. So now we are getting look at the p value. The p value is 1.7 10 to the power – 11 it is very low value. So we have to reject our null hypothesis. You can look at this our calculated chi square value is 74.

The table chi square value is 24 so 74 is larger than 24 so we have to reject our null hypothesis then we are concluding that the given dataset is not following uniform distribution.

(Refer Slide Time: 21:26)

```

In [21]: mean = np.mean(A)
mean
Out[21]: 71.0

In [22]: std = np.std(A)
std
Out[22]: 18.22635454499845

In [23]: x = 1/6 * np.arange(6)

In [24]: for j in range(1,6):
    Prob_intervals = [scipy.stats.norm.pdf(j*x, mean, std)]
    print(Prob_intervals)
[53.36742154275226]
[63.14941153988116]
[71.0]
[78.8958640918894]
[88.6356845424763]

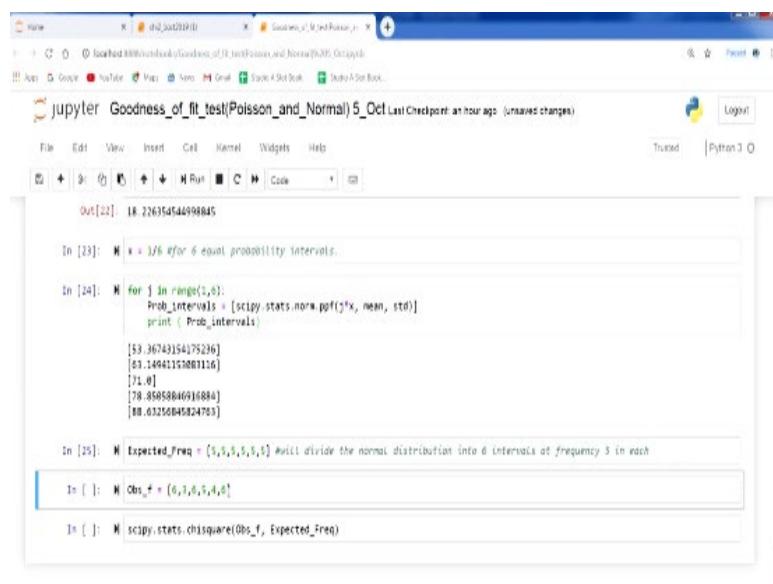
In [ ]: Expected_Freq = {1,1,1,1,1,1} #will divide the normal distribution into 6 intervals of frequency 5 in each

```

After that you will see the second example there are some dataset is given we are going to test whether this dataset follow normal distribution or not. So what I am running this dataset. First I am finding the mean the mean of the given dataset is 71 and the standard deviation of the given dataset is 18.23. So what is a null hypothesis the given dataset whose mean is 71 and standard deviation 18.22 follow normal distribution.

Alternative hypothesis is it is not following normal distribution then the given dataset the x value we have to divide by 6 because we need to have minimum 5 expected frequency not observed frequency, expected frequency in each interval. So the given dataset is divided into 6 so that because 30 divided by 6 I will get 5 expected frequency in each interval. Now in the range of 1, 6 I am going to get the different intervals. So one interval is 0 to 53.36 another interval is 53.36 to 63.14 another interval is 63.14 to 71 next one 71 to 78.85 the next interval is 88.63 and above. So this was our interval.

(Refer Slide Time: 22:54)



The screenshot shows a Jupyter Notebook window titled "jupyter Goodness_of_fit_test(Poisson_and_Normal) 5_Out". The notebook has several cells:

- In [22]:** `18.2263546998845`
- In [23]:** `# x = 1/6 for 6 equal probability intervals.`
- In [24]:**

```
# for j in range(1,6):
    Prob_intervals = [scipy.stats.norm.pdf(j*x, mean, std)]
    print (Prob_intervals)
```

```
[53.36743154175236]
[53.14941115880116]
[71.0]
[78.8595846916884]
[10.63250484824763]
```
- In [25]:** `# Expected_Freq = [5,5,5,5,5,5] #will divide the normal distribution into 6 intervals of frequency 5 in each`
- In []:** `Obs_f = [6,1,5,5,4,5]`
- In []:** `scipy.stats.chisquare(Obs_f, Expected_Freq)`

So, how we got this interval now we got the different intervals. Our expected frequency is 5,5,5 I am running this because we have divided by 6 and our observed frequency. So from the given dataset in the range of 0 to 53.67 we have to count it how many dataset is there, there are 6 dataset. In the interval 53.36 to 63.14 there are 3 dataset. In the interval 63.14 to 71 there are 6 dataset and so on. We have to manually count it how many dataset is appearing in this interval that is our observed frequency. So now we got the expected frequency and observed frequency.

(Refer Slide Time: 23:39)

The screenshot shows a Jupyter Notebook window with several code cells. Cell In [14] contains a loop to calculate probability density function (PDF) values for a normal distribution across six intervals. Cell In [25] defines a variable `Expected_Freq` as a list of 5s. Cell In [26] defines `Obs_f` as a list of observed frequencies. Cell In [27] performs a chi-square test using `scipy.stats.chisquare`, resulting in a `Power_DivergenceResult` object with a statistic of 1.1999999999999999 and a p-value of 0.9012491445012737.

```

In [14]: for j in range(1,6):
    Prob_intervals = [scipy.stats.norm.pdf(j*x, mean, std)]
    print (Prob_intervals)

[53.36743154575216]
[63.14941153080116]
[71.0]
[78.8595846016884]
[88.63256045424763]

In [25]: Expected_Freq = [5,5,5,5,5,5] #will divide the normal distribution into 6 intervals of frequency 5 in each

In [26]: Obs_f = [6,3,6,5,4,6]

In [27]: scipy.stats.chisquare(Obs_f, Expected_Freq)
Out[27]: Power_DivergenceResult(statistic=1.1999999999999999, pvalue=0.9012491445012737)

```

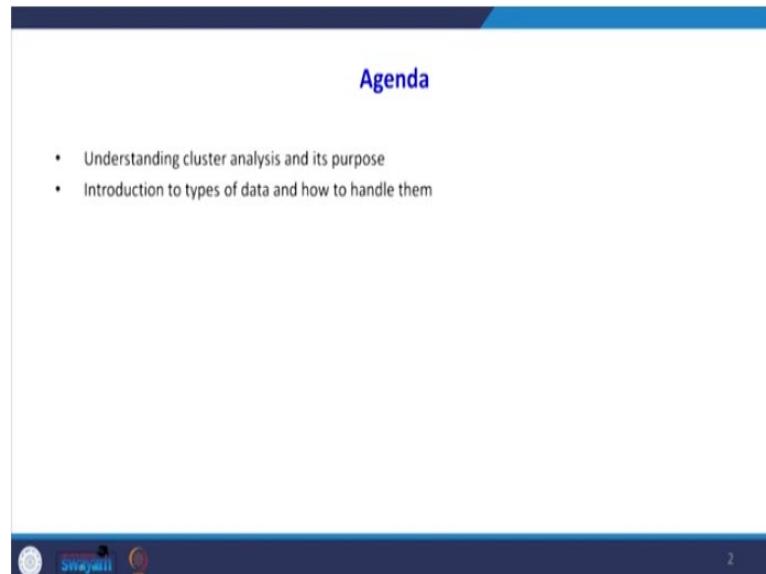
Then when we do the chi square test so we are getting p value 0.90 that is bigger than our alpha value. So we have to accept null hypothesis and we are concluding that the given dataset follow normal distribution. In this lecture, first I have explained a demo for testing goodness of fit for a Poisson distribution then I have explained the theory behind how to test goodness of fit for a uniform distribution normal distribution that is some dataset is given.

How to test the given dataset follow uniform distribution or normal distribution then I have done the python code with the help of python demo I have explained how to test the uniform and normal distribution for goodness of fit because the testing another important point you have to remember all the random numbers follow uniform distribution. Sometime you may ask to say some random numbers.

And you have to test whether the numbers are really random or not. For that purpose, we have to test whether the given dataset it is following uniform distribution or not. If certain numbers following uniform distribution we can conclude that, that numbers are random numbers.

Lecture – 49
Cluster Analysis: Introduction - 1

(Refer Slide Time: 00:36)



The slide has a blue header bar and a blue footer bar. In the footer bar, there are three small icons: a blue one, a red one with the word 'Swajam', and a yellow one. The number '2' is also visible in the bottom right corner of the footer.

Agenda

- Understanding cluster analysis and its purpose
- Introduction to types of data and how to handle them

Dear students today we are entering to a new topic that is a cluster analysis. The cluster analysis is mostly widely used data mining techniques. It is a very important topic, so the agenda for today class is understanding cluster analysis and its purpose. Then introduction to types of data and how to handle them because the clustering techniques will vary with respect to what kind of data nature of the data is; suppose the nature of the data is continuous data or interval data, there will be a different algorithm for that. If the data is nominal data, there will be a different algorithm for clustering.

(Refer Slide Time: 01:06)

Cluster analysis

- The classification of similar objects into groups is an important human activity, this is part of the learning process
- i.e. A child learns to distinguish between cats and dogs, between tables and chairs, between men and women, by means of continuously improving subconscious classification schemes
- This explains why cluster analysis is often considered as a branch of pattern recognition and artificial intelligence



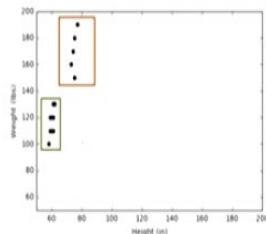
First, we will see what is cluster analysis? Cluster analysis is the art of finding groups in data. In cluster analysis basically one wants to form groups in such a way that objects in the same groups are similar to each other whereas the object in different groups are dissimilar as possible. When you look at this picture, they are different clusters say group of dogs, group of cat, group of chairs, group of tables. This is an example of cluster analysis.

The classification of similar objects into groups an important human activity. This is part of learning process. A child learns to distinguish between cats and dogs, between tables and chairs, between men and women by means of continuously improving subconscious classification schemes. What is the meaning of this point is a child unknowingly is able to classify different objects and able to group cluster different objects which are similar in nature.

(Refer Slide Time: 02:17)

Example

- Lets illustrate with the help of an example:
- It is a plot of twelve objects, on which two variables were measured. For instance, the weight of an object might be displayed on the vertical axis and its height on the horizontal one



5

We will explain the concept of cluster analysis with the help of an example. This is a plot of 12 objects on which two variables were measured. For instance, the weight of an object might be displayed on the vertical axis and its height on the horizontal axis when you plot it here it is clearly visible that you are able to form two clusters with respect to height you see that these are the different weight with respect to another height to this is the 60 is one type of height and second is 80 is another type of height we are able to cluster it.

(Refer Slide Time: 02:56)

Example

- Because this example contains only two variables, we can investigate it by merely looking at the plot
- In this small data set there are clearly two distinct groups of objects
- Such groups are called clusters, and to discover them is the aim of cluster analysis

6

Because this is contains only two variable. We can investigate it by merely looking at the plot. In this small data set that are clearly two distinct group of objects. Such groups are called clusters and to discover them is the aim of cluster analysis, so this is the purpose of our class. What is

going to be there in coming classes. We will be having different types of data. We are going to cluster that different data into different groups.

(Refer Slide Time: 03:32)

Cluster and discriminant analysis

- Cluster Analysis is an unsupervised classification technique in the sense that it is applied to a dataset where patterns want to be discovered (i.e. groups of individuals or variables want to be found)
- No prior knowledge is needed for this grouping, and it is sensitive to several decisions that have to be taken (similarity/dissimilarity measures, clustering method,...)
- Discriminant Analysis (DA) is a statistical technique used to build a prediction model that is used to classify objects from a dataset depending on the features observed on them. In this case, the dependent variable is the grouping variable, which identifies to which group and object belongs
This grouping variable should be known at the beginning, for the function to be built up. Sometimes DA is considered as a Supervised tool, as there is a previous known classification for the elements of the dataset



Many time the students may have doubt what is the difference between cluster and discriminant analysis. Cluster analysis is an unsupervised classification technique in the sense that it is applied to a data set where patterns want to be discovered. That is the group of individuals or variables wanted to be found. Why we are calling it this unsupervised learning because we may not know which variable will go to which cluster, we are not knowing also that how many clusters we are going to form it.

The second point in a cluster analysis, no prior knowledge is needed for this grouping. I need to sensitive to several decisions that have to be taken. Some of the variables are similarity dissimilarity measures clustering methods. Whereas discriminant analysis is a statistical techniques used to build a prediction model that is used to classify objects from your data set depending on the futures observed on them.

In this case, the dependent variable is grouping variable which identifies to which group or object belongs. This grouping variable should be known at the beginning for a function to be built up sometime discriminant analysis is considers supervised tool because as there is a previous known classification for the element of the dataset.

(Refer Slide Time: 05:03)

Cluster analysis and discriminant analysis

- Cluster analysis can be used not only to identify a structure already present in the data, but also to impose a structure on a more or less homogeneous data set that has to be split up in a "fair" way, for instance when dividing a country into telephone areas
- Cluster analysis is quite different from discriminant analysis in that it actually establishes the groups, whereas discriminant analysis assigns objects to groups that were defined in advance



Further we will continue the difference between a cluster analysis and discriminant analysis. Cluster analysis can be used not only to identify the structure already present in the data, but also to impose structure on a more or less homogeneous data set that has to be split up in a fair way. For instance, when dividing a country into telephone areas. See this is a country for example see this example this is classified into different telephone areas.

Cluster analysis is quite different from discriminant analysis in that it actually establishes the groups whereas discriminant analysis assigns object to groups that were defined in advance. That is a major difference. What does that mean? The discriminant analysis assigns object to the group that were defined in advance. But in cluster analysis it is not the case and what will happen as I told you in the beginning, the clustering analysis and corresponding algorithm depending upon what kind of data.

(Refer Slide Time: 06:07)

Types of data and how to handle them

- Let us take an example, there are n objects to be clustered, which may be persons, flowers, words, countries, or anything
- Clustering algorithms typically operate on either of two input structures:
 - The first represents the objects by means of p measurements or attributes, such as height, weight, sex, color, and so on
 - These measurements can be arranged in an n -by- p matrix, where the rows correspond to the objects and the columns to the attributes



Types of data and how to handle them for cluster analysis, as I told you in the beginning of the class, the types of data is an important point has to be taken care while doing cluster analysis because for different types of data there is a different type of clustering algorithms. Let us take an example that are n objects to be clustered, which may be persons, flowers, birds, countries or anything.

Clustering algorithm typically operate on either of two input structure. The first represents the object by means of p measurement or attributes such as height, weight, sex color and so on. These measurements can be arranged in a n -by- p matrix whereas the row corresponds to the objects and the column corresponds to the attributes.

(Refer Slide Time: 07:01)

Example

Attributes

Objects

	Price	Quality	Time
Like	A	B	B
Intermediate	B	A	A
Need	C	C	C



10

You see this case the objects are in the rows Like, Intermediate, Need the attributes, Price, Quality, Times are in the columns this is one kind of input.

(Refer Slide Time: 07:11)

Types of data and how to handle them

- The second structure is a collection of proximities that must be available for all pairs of objects
- These proximities make up an n-by-n table, which is called a one-mode matrix because the row and column entities are the same set of objects
- one shall consider two types of proximities, namely dissimilarities (which measure how far away two objects are from each other) and similarities (which measure how much they resemble each other)

	A	B	C
A	1		
B		1	
C			1



11

The second structure is a collection of proximities that must be available for all pairs of objects. These proximities makeup an n by n table, which is called one mode matrix because the row and the column entities are the same set of objects. One shall consider two type of proximities, namely dissimilarities which measures how far away two objects are from each other and similarities which measures how much they are resemble each other okay.

Now assume that there are some variable A, B, C, so the A and B it is written this way. You see that A and A, B and B, C and C is 1 because the same one. So we can write between A and B how much is the similarity otherwise between A and B how much is dissimilarity.

(Refer Slide Time: 08:18)

Type of data

- **Interval-Scaled Variables**
- In this situation the n objects are characterized by p continuous measurements
- These values are positive or negative real numbers, such as height, weight, temperature, age, cost, ..., which follow a linear scale
- For instance, the time interval between 1900 and 1910 was equal in length to that between 1960 and 1970.

Time scale in years

Smashit

12

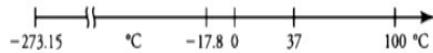
Now let us see what is the interval scaled variables. In this situation the n objects are characterized by p continuous measurement. These values are positive or negative real numbers such as height, weight, temperature, age, cost which follow a linear scale. For instance, the time interval between 1900 and 1910 was equal in length to that between 1960 and 1970. So this is an example of interval scales we have studied in the beginning of the lecture.

We have classified data in to different four categories, nominal, ordinal, interval, ratio. So, the example of interval is year. If it is a year, what happened? We can add some numbers, we can subtract some numbers. Similarly, you see that between 19 the interval will be same between 1900 to 1910 and 1960 to 1970 the difference is same because we can add it, we can subtract it, but we cannot multiply.

(Refer Slide Time: 09:18)

Type of data

- Also, it takes the same amount of energy to heat an object of -16.4°C to -12.4°C as to increase it from 35.2°C to 39.2°C
- In general it is required that intervals keep the same importance throughout the scale



13

Similarly the another example for interval data is our temperature. Also it takes the same amount of energy to heat an object of -14.4 degrees Celsius to -12.4 degrees Celsius as to increase it from 35.2 degrees Celsius to 39.2 degree Celsius. What I am saying is that the Fahrenheit temperature scale also an example of interval scale because there would not be any absolute 0 but we can add, we can subtract. In general it is required that intervals keeps the same importance throughout the scale.

(Refer Slide Time: 09:55)

Interval-Scaled Variables

- These measurements can be organized in an n -by- p matrix, where the rows correspond to the objects (or cases) and the columns correspond to the variables.
- When the f^{th} measurement of the i^{th} object is denoted by x_{if} (where $i = 1, \dots, n$ and $f = 1, \dots, p$) this matrix looks like:

$$\begin{matrix} & \downarrow & \downarrow & \downarrow & J \\ & p \text{ variables} & & & \\ \text{n objects} & \left[\begin{array}{cccc} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \\ \vdots & & \vdots & & \vdots \\ x_{s1} & \cdots & x_{sf} & \cdots & x_{sp} \end{array} \right] \end{matrix}$$

14

Interval scale scaled variables. These measurements can be organized in an n -by- p matrix where the rows corresponds to the objects and the column corresponds to the variables. So where, the f^{th} measurement of the i^{th} object is denoted by x_{if} where i is 1 to n , f is 1 to p . So, here in row

we have mentioned objects in column we have mentioned variables. The another name for object in cases the variable may be different variables

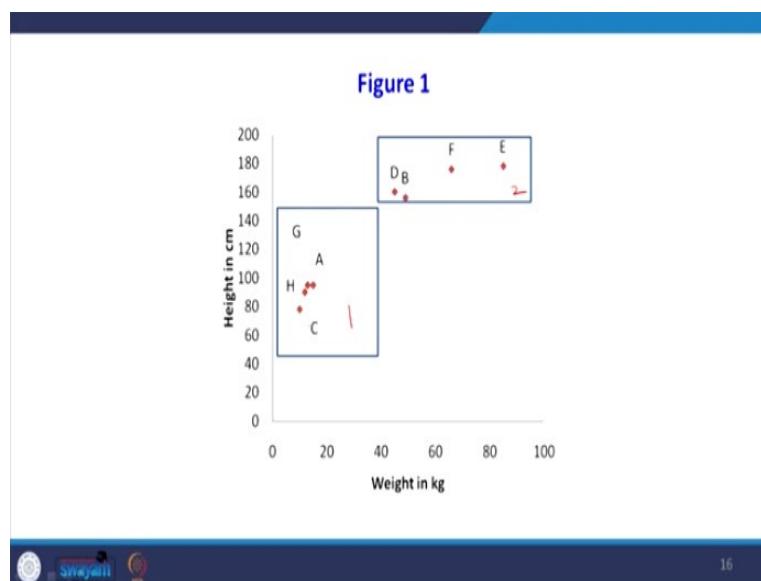
(Refer Slide Time: 10:33)

Interval-Scaled Variables		
Person	Weight(Kg)	Height(cm)
A	15	95
B	49	156
C	13	95
D	45	160
E	85	178
F	66	176
G	12	90
H	10	78

Table :1

For example take 8 people for example what is happening in rows we have n objects in column we have 2 variables one variable is height another variable is weight. Take 8 people the weight in kilogram and the height in centimeter is given in the table. In this situation n = 8 because 8 rows are there p = 2 because 2 variable is there.

(Refer Slide Time: 11:01)



If I plot that you see that weight is taken in kg height in centimeter. When you plot it, it is forming two similar objects. We can group into similar objects into two category one is cluster 1

we can call it is cluster 1 and cluster 2. In cluster 1 for example your row C H A G will occur in cluster 2 row D B F E will occur. This is a one way of clustering.

(Refer Slide Time: 11:35)

Interval-Scaled Variables

- The units on the vertical axis are drawn to the same size as those under horizontal axis even though they represent different physical concepts
- The plot contains two obvious clusters, which can in this case be interpreted easily: the one consists of small children and the other of adults
- However, other variables might have led to completely different clustering
- For instance, measuring the concentration of certain natural hormones might have yielded a clear cut partition into different male and female persons

The units on the vertical axis are drawn to the same size as those under horizontal axis even though they represent different physical concepts. The plot contains two obvious clusters which can in this case be interpreted easily. The one consists of small children other of adult. However other variables might have led to completely different clustering. For instance, measuring the concentration of certain natural hormones might have yielded a clear cut partition into different male and female persons. In this one since we are taken 2 variable one is weight and height instead of weight and height if you take some other variables that may bring some other type of clustering.

(Refer Slide Time: 12:21)

Interval-Scaled Variables

- Let us now consider the effect of changing measurement units.
- If weight and height of the subjects had been expressed in pounds and inches, the results would have looked quite different.
- A pound equals 0.4536 kg and an inch is 2.54 cm
- Therefore, Table 2 contains larger numbers in the column of weights and smaller numbers in the column of heights. Figure 2

Person	Weight(lb)	Height(in)
A	33.1	37.4
B	108	61.4
C	28.7	37.4
D	99.2	63
E	187.4	70
F	145.5	69.3
G	26.5	35.4
H	22	30.7

Table :2



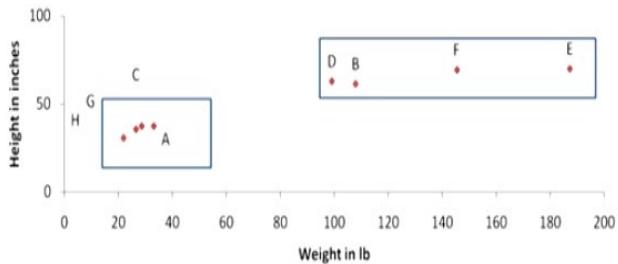
18

Let us now consider the effect of changing the measurement unit. So previously we had the measurement unit and you look at this one the weight is in kg and height is in centimeter. Now if you change the unit now, the weights going to be in pound and height is going to be in inch. Now for this kind of dataset, let us see how this unit of this data is going to affect our clustering technique.

Let us now consider the effect of changing measurement units. If weight and height of the subject had been expressed in pounds and inches the result would have looked quite different. A pound equals 0.453 kg and an inch is 2.5 centimeter. Therefore, table 2 contains larger number in column of weight because we are converted into pounds and smaller number in the column of heights the heights become very smaller.

(Refer Slide Time: 13:18)

Figure 2



19

Let us see the new cluster now what happened? Now the height has increased, the weight has increased. Now the clustering pattern is completely changed. So what point I am saying here is that the unit of the data may bring out different clusters.

(Refer Slide Time: 13:36)

Interpretation

- Although plotting essentially the same data as Figure 1, Figure 2 looks much flatter
- In this figure, the relative importance of the variable "weight" is much larger than in Figure 1
- As a consequence, the two clusters are not as nicely separated as in Figure 1 because in this particular example the height of a person gives a better indication of adulthood than his or her weight. If height had been expressed in feet ($1 \text{ ft} = 30.48 \text{ cm}$), the plot would become flatter still and the variable "weight" would be rather dominant
- In some applications, changing the measurement units may even lead one to see a very different clustering structure

20

So what is interpretation, although plotting essentially the same data as figure 1, figure 2 looks much flatter. In this figure the relative importance of the variable weight is much larger than the figure 1. As a consequence, the two clusters are not as nicely separated as in figure 1 because in this particular example, the height of the person gives a better indication of adulthood then his or her weight.

If height had been expressed in feet because 1 feet = 30.48 centimeter the plot would become flatter still and the variable weight would be rather dominant. In some applications changing the measurement units that is an important point, may even lead to one to see a very different clustering structures. The point what I am trying to say here is that changing the measurement unit may provide different type of clustering structure.

(Refer Slide Time: 14:37)

Standardizing the data

- To avoid this dependence on the choice of measurement units, one has the option of standardizing the data
- This converts the original measurements to unitless variables
- First one calculates the mean value of variable f , given by:

$$m_f = \frac{1}{n} (x_{1f} + x_{2f} + \dots + x_{nf})$$

for each $f = 1, \dots, p$

To avoid this because different units are providing different clustering structure one way to avoid this problem is standardizing the data. Now let us see how to standardize the data to avoid this dependence on the choice of measurement units one has the option of standardizing the data. This converts the original measurement to unitless variable. First one calculates the mean value of f given by we know that the mean is $1 / n$ sum of all the values divided by m number of variables that is m_f .

(Refer Slide Time: 15:18)

Standardizing the data

- Then one computes a measure of the dispersion or "spread" of this f^{th} variable
- Generally, we use the standard deviation for this purpose

$$\text{std}_f = \sqrt{\frac{1}{n-1} \left((x_{1f} - m_f)^2 + (x_{2f} - m_f)^2 + \dots + (x_{nf} - m_f)^2 \right)}$$

22

Then one computes a measure of dispersion or spread of this f^{th} variable. Generally we use the standard deviation what is a standard deviation x_1 that variable minus mean whole square divided by, second variable minus whole square, up to f^{th} variable - m_f whole square divided by $n - 1$ that is a standard deviation. This is one way of standardizing the data.

(Refer Slide Time: 15:44)

Standardizing the data

- However, this measure is affected very much by the presence of outlying values
- For instance, suppose that one of the x_{if} has been wrongly recorded, so that it is much too large
- In this case std_f will be unduly inflated, because $x_{if} - m_i$ is squared
- Hartigan (1975, p. 299) notes that one needs a dispersion measure that is not too sensitive to outliers
- Therefore, we will use the mean absolute deviation, where the contribution of each measurement x_{if} is proportional to the absolute value $|x_{if} - m_i|$

$$s_f = \frac{1}{n} \left(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f| \right)$$

23

However this measure is affected very much by the presence of outlying values. The problem with the standardization is that if there are extremely large values or extremely low values that is affecting the process of standardization. For instance, suppose that one of the x_{if} has been wrongly recorded so that it is much too large. In this case, the standard deviation will be unduly inflated because we are squaring $x_{if} - m_i$ is squared .

So Hartigan in the year 1975 notes that one needs a dispersion measure that is not too sensitive to outliers. Therefore we will use the mean absolute deviation we generally this term generally we call it as MAD mean absolute deviation where the contribution of each measurement x_{if} is proportional to the absolute value of modulus value of $x_{if} - m_f$. So instead of squaring, we are going to take to find the standardize we are going to take only the mean absolute deviation.

The advantage of taking mean absolute deviation is that if any out layer is there that its effect is dampened. That is why instead of going for standard deviation, we should go for mean absolute deviation. That is $x_f = 1/n (|x_1f - m_f| + |x_2f - m_f| \text{ modulus and so on})$.

(Refer Slide Time: 17:21)

Standardizing the data

- Let us assume that s_f is nonzero (otherwise variable f is constant over all objects and must be removed)
- Then the standardized measurements are defined by and sometimes called z-scores
- They are unitless because both the numerator and the denominator are expressed in the same units
- By construction, the z_{if} have mean value zero and their mean absolute deviation is equal to 1

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$
(0, 1)

24

Let us assume that S_f is nonzero because the standardized value should be non-zero otherwise the variable f is constant over all objects and must be removed. Then the standardized measurement are defined by and sometimes called z-scores. The another name for standardization is z-score. They are unitless because both the numerator, because numerator also deviation the denominators are also deviation.

They are unitless because the numerator, the denominator are expressed in the same units. By construction z_{if} have mean 0 and then absolute deviation is equal to 1. So what is happening in the property of this where standardized data is mean should be 0 and the variance or standard division should be 1.

(Refer Slide Time: 18:10)

Standardizing the data

- When applying standardization, one forgets about the original data and uses the new data matrix in all subsequent computations

$$\text{objects} \begin{bmatrix} z_{11} & \cdots & z_{1f} & \cdots & z_{1p} \\ \vdots & & \vdots & & \vdots \\ z_{i1} & \cdots & z_{if} & \cdots & z_{ip} \\ \vdots & & \vdots & & \vdots \\ z_{n1} & \cdots & z_{nf} & \cdots & z_{np} \end{bmatrix}$$

25

When applying standardization one forgets about the original data and uses the new data matrix in all subsequent computations. What happened initially in row there was object in column there was variables. There was x_{if} variable was there. So after standardization, that will become z_{11}, z_{12} up to z_{1p} . Now this data that is data which are standardized data will be taken for further analysis.

(Refer Slide Time: 18:40)

Detecting outlier

- The advantage of using s_i rather than std_p in the denominator of z-score formula is that s_i will not be blown up so much in the case of an outlying x_{ip} , and hence the corresponding z_{if} will still be noticeable so the i^{th} object can be recognized as an outlier by the clustering algorithm, which will typically put it in a separate cluster

26

Detecting outlier the advantage of using S_f rather than standardized value of f in the denominator of z-score formula is that S_f will not be blown up so much in the case of an outlying x_{if} and hence the corresponding z_{if} will still be noticeable. So the i^{th} object can be recognized as an

outlier by the clustering algorithm, which will typically put it in a separate cluster. So the purpose of using the z-score is it will not blow up so much in the case of any outlier in the dataset.

(Refer Slide Time: 19:18)

Standardizing the data

- The preceding description might convey the impression that standardization would be beneficial in all situations.
- However, it is merely an option that may or may not be useful in a given application
- Sometimes the variables have an absolute meaning, and should not be standardized
- For instance, it may happen that several variables are expressed in the same units, so they should not be divided by different s_i
- Often standardization dampens a clustering structure by reducing the large effects because the variables with a big contribution are divided by a large s_i

Standardizing the data the preceding description might convey the impression that the standardization would be beneficial in all situations. However, it is merely an option that may or may not be useful in a given application. Sometimes the variables have an absolute meaning and should not be standardized. What is the point here is that the variable already in the absolute term, it should not be standardized.

For instance, it may happen that several variables are expressed in the same units, so they should not be divided by different S_f . Because all the variables are in the same units you need not go for standardization. Often standardization dampens a clustering structure by reducing the large effect because the variables with the big contribution are divided by a large S_f . S_f is standardized value.

In this lecture we have covered the purpose of clustering analysis. Then I have explained the difference between clustering analysis and discriminant analysis. Then I have explained how the different types of data will affect our clustering structure. In the different types of data we have taken only the interval data and how to handle them for doing cluster analysis.

Then I have started why we have to do the standardization because if the different variables are in different units, you may get different kinds of clustering structures. To overcome that we have to go for standardization. The next class I will explain that the standardization also not applicable for all kind of data. Sometime it will mislead it may provide a different type of clustering that I will explain with the help of an example in the next class. Thank you.

Department of Computer Science and Engineering

Indian Institute of Technology - Roorkee

Lecture – 50
Clustering Analysis: Part II

In my previous lecture, we have started about introduction to cluster analysis, and I have explained how to handle interval types of data. Then I have started about the importance of standardization. In this lecture we will see that what is the effect of standardization because sometime standardization may mislead your clustering structure and I will explain different types of distances computation between the objects.

(Refer Slide Time: 00:56)

Agenda

- Explain effect of standardization(with help of an example)
- Different types of distances computation between the objects
- Handling missing data

Because for different types of data set, there are different ways to compute the distances, so that I will explain the many time when we collect the data. It is not necessary that we will collect all the data some time there may be a missing data. If the data is missed how to hold to handle that, that also will cover in this lecture.

(Refer Slide Time: 01:13)

Example

- Lets take four persons A, B, C, D with following age and height:

Person	Age (yr)	Height (cm)
A	35	190
B	40	190
C	35	160
D	40	160

TABLE: 1

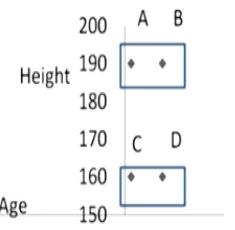


FIGURE: 1

Finding Groups in Data: An Introduction to Cluster Analysis
Author(s): Leonard Kaufman, Peter J. Rousseeuw
March 1990, John Wiley & Sons, Inc.



3

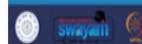
Now let us see the effect of standardization I have taken one simple problem with a numerical example. This problem is taken from this book Finding Groups in Data: An Introduction to Cluster Analysis by Leonard Kaufman and Peter Rousseeuw; it is a John Wiley publishers. There are 4 persons and your age yet in terms of year and height in terms of centimeter is given. Suppose if you take age on horizontal axis and height on vertical axis, you can mark this all four persons A, B, C, D.

So what you were able to understand that is a distinct cluster is there. Because A, B is one group one cluster C, D in another cluster. Now the same data let us standardize after standardizing again we will go for clustering let us see how it appears.

(Refer Slide Time: 02:08)

Example

- In Figure 1 we can see two distinct clusters
- Let us standardize the data of Table 1
- The mean age equals $m_1 = 37.5$ and the mean absolute deviation of the first variable works out to be $s_1 = (2.5 + 2.5 + 2.5 + 2.5)/4 = 2.5$
- Therefore, standardization converts age 40 to +1 ($(40-37.5)/2.5 = 1$) and age 35 ($(35 - 37.5)/2.5 = -1$) to -1
- Analogously, $m_2 = 175$ cm and $s_2 = (15 + 15 + 15 + 15)/4 = 15$ cm, so 190 cm is standardized to +1 and 160 cm to -1



4

In figure 1 we can see the distinct clusters, let us standardize the data of table 1. For standardizing we should know the mean and standard deviation, standard deviation otherwise mean absolute deviation. So that mean of age equals to $m_1 = 37.5$ just by adding all the ages and divided by the number of data set and the mean absolute deviation is not standard deviation it is mean absolute deviation of the first variable works out to be $S_1 = 2.5$.

How we are finding mean absolute deviation that variable minus mean for example $35 - 37.5$ for second variable $40 - 37.5$ we have to take only the positive value. There are four data set, so the mean absolute deviation is 2.5. Therefore, the standardization convert 40 to +1 how we got to 40 is converted standardized to 1 we know that this is $(x - \mu)$ divided by S. So x is 40 mu that is m is 37.5 divided by mean absolute deviation 2.5, = 1.

And same way age 35 is standardized to -1 how we got the -1, 35 - mean divided by mean absolute deviation. So it is - 2.5 divided by 2.5 it is - 1 the same way for the variable m_2 the mean is 175 and mean absolute deviation for variable 2 is 15. So each variable in the second column also standardized for example 190 centimeter is standardized to + 1 and same way 160 centimeters is standardized to -1.

(Refer Slide Time: 03:53)

Example

- The resulting data matrix, which is unitless, is given in Table 2
- Note that the new averages are zero and that the mean deviations equal 1

• Table 2

Person	Variable 1	Variable 2
A	1	1
B	-1	1
C	1	-1
D	-1	-1

- Even when the data are converted to very strange units standardization will always yield the same numbers



5

The result data matrix which is unitless because below standardized is given in the table 2. Note that the new averages are 0 and the mean deviations equal to 1. So this table 2 shows that these standardized. Table for each variable is variable 1 and variable 2. Even when the data are converted into various strange units standardization will always yield the same numbers that is the advantage of standardization.

(Refer Slide Time: 04:25)

Example

- Plotting the values of Table 2 in Figure 2 does not give a very exciting result
- Figure 2 shows no clustering structure because the four points lie at the vertices of a square
- One could say that there are four clusters, each consisting of a single point, or that there is only one big cluster containing four points
- Here standardizing is no solution

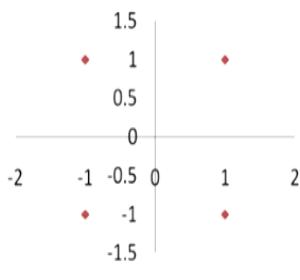


FIGURE: 2



6

Now plotting the values of table 2 in the figure 2 does not give any very exciting result. So what do you have done? In the previous table we have the standardized values for both the variables. So when you plot it there are 4 points are appearing. So this points is not giving any useful result so figure 2 shows no clustering structure because 4 points lay out the vertices of a square. One

could say that there are 4 clusters; each consisting of single point are that there is only one big cluster containing 4 points.

Here standardization is no solution. So what we have seen many times when you go for standardization, the standardization may not give the useful result that is what this example shows.

(Refer Slide Time: 05:16)

Choice of measurement (Units)- Merits and demerits

- The choice of measurement units gives rise to relative weights of the variables
- Expressing a variable in smaller units will lead to a larger range for that variable, which will then have a large effect on the resulting structure
- On the other hand, by standardizing one attempts to give all variables an equal weight, in the hope of achieving objectivity
- As such, it may be used by a practitioner who possesses no prior knowledge



7

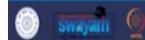
Now let us look at the choice of measurements. Here the measurement means that units of that variable. What is the merits and demerits? The choice of measurement units gives rise to relative weight of variables, expressing a variable in smaller units will lead to large range for that variable, which will then have a large effect on the resulting structure. So what will happen if variable is in smaller units? So, that will give a larger effect in the; your clustering result.

On the other hand, by standardizing one attempts to give all variables an equal weight in the hope that achieving objectivity. As such it may be used for practitioners who possesses no prior knowledge. So the benefit of standardization is that anybody those who are not having any prior knowledge about the problem also can do with the help of standardized variables. They can do the cluster analysis because there is a unitless.

(Refer Slide Time: 06:18)

Choice of measurement- Merits and demerits

- However, it may well be that some variables are intrinsically more important than others in a particular application, and then the assignment of weights should be based on subject-matter knowledge
- On the other hand, there have been attempts to devise clustering techniques that are independent of the scale of the variables



8

However, it may well be that some variables are intrinsically more important than others in a particular application and then the assignment of weight should be based on the subject matter knowledge. Every time because standardization is giving equal weight some time some variables are more important. So for that variable with the help of experts, we can give a higher weightage for that variable.

On the other hand, there have been attempts to devise clustering techniques that are independent of scale of the variables. There are many techniques people are trying to come with a different clustering model.

(Refer Slide Time: 06:55)

Distances computation between the objects

- The next step is to compute distances between the objects, in order to quantify their degree of dissimilarity
- It is necessary to have a distance for each pair of objects i and j.
- The most popular choice is the Euclidean distance:

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}$$

- When the data are being standardized, one has to replace all x by z in this expression
- This Formula corresponds to the true geometrical distance between the points with coordinates (x_{i1}, \dots, x_{ip}) and (x_{j1}, \dots, x_{jp})



9

Distances computation between objects. The next step is to compute distances between the objects in order to quantify their degree of dissimilarity. It is necessary to have a distance for each pair of objects i and j . The most popular choice is the Euclidean distance. What is this Euclidean distance? The distance between variable $i, j = (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2$ up to $(x_{ip} - x_{jp})^2$.

When the data are being standardized one has to replace all x by z in this expression if you are standardizing instead of x you have to use z . This formula corresponds to the true geometrical distance between points with the coordinates x_{i1} up to x_{ip} and x_{j1} up to x_{jp} .

(Refer Slide Time: 07:55)

Example

- let us consider the special case with $p = 2$ (Figure 3)
- Figure shows two points with coordinates (x_{i1}, x_{i2}) and (x_{j1}, x_{j2})
- It is clear that the actual distance between objects i and j is given by the length of the hypotenuse of the triangle, yielding expression in previous slide by virtue of Pythagoras' theorem

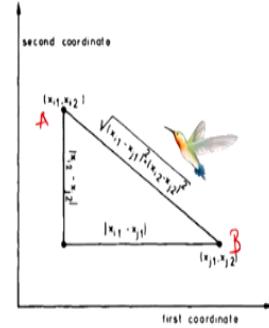


Figure 3: Illustration of the Euclidean distance formula



10

See the Euclidean distance suppose if you want to move from point A to B see this is point A and B let us find out the concept behind the Euclidean distance. Suppose if we want to move if you want to point A to B you can directly you can fly from one point to because the birds will fly from A to B. So that distances called Euclidean distance. Let us consider the special case with $p = 2$ where there are only two variables.

Figure shows two points with the coordinates x_{i1}, x_{i2} and x_{j1}, x_{j2} . It is clear that the actual distance between objects i and j is given by the length of the hypotenuse of the triangle yielding expression in previous slide by virtue of Pythagoras theorem. So this formula is nothing but the

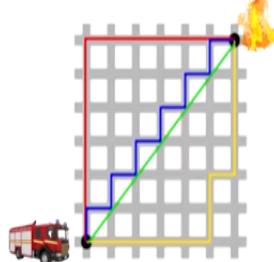
hypotenuse. So this is as per the Pythagoras theorem so square of adjacent side and square of opposite side equal to square of hypotenuse.

(Refer Slide Time: 09:02)

Distances computation between the objects

- Another well-known metric is the city block or Manhattan distance, defined by:

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$



11

Let us go to the next distance measures that is Manhattan distance. It is another well-known metric is the city block or Manhattan distance. It is given by modulus value of $x_{i1} - x_{j1} + x_{i2} - x_{j2}$ modulus value only the positive values up to $x_{ip} - x_{jp}$. Suppose this is city map if you want to move from point A to B right. There are two ways one way is directly you can suppose we are if you are a bird or you are move if you want to go A to B you can fly. Otherwise, the flight goes from point A to point B.

But if there is a fire suppose a fire engine it has to move. It has to follow a rectangular distance. So because there are different streets, so this distance is nothing but your Manhattan distance. You see that the distance or Manhattan distance will be larger than the Euclidean distance the green one is Euclidean distance. The blue one is nothing but the Manhattan distance.

(Refer Slide Time: 10:10)

Interpretation

- Suppose you live in a city where the streets are all north-south or east-west, and hence perpendicular to each other
- Let Figure 3 be part of a street map of such a city, where the streets are portrayed as vertical and horizontal lines



12

Let us interpret the meaning of Manhattan distance. Suppose you live in a city where the streets are all north-south or east-west and hence perpendicular to each other. Let figure 3 be the part of street map of such a city where the streets are portrayed as a vertical and horizontal lines. So if you want to move from point A to point B, you cannot directly you cannot go by shortest path you have to take a rectangular distance. Another name for this Manhattan distance is rectilinear distance.

(Refer Slide Time: 10:41)

Interpretation

- Then the actual distance you would have to travel by car to get from location i to location j would total $|x_{i1} - x_{j1}| + |x_{i2} - x_{j2}|$
- This would be the shortest length among all possible paths from i to j
- Only a bird could fly straight from point i to point j, thereby covering the Euclidean distance between these points



13

Then the actual distance you would have to travel by a car or fire engine to get from location i to location j would be $x_{i1} - x_{j1}$ modulus value + $x_{i2} - x_{j2}$ modulus value. This would be the shortest length among all possible paths from i to j. Only a bird could fly straight from point i to j

thereby covering Euclidean distance between these points. So the example of the bird, which covers point A to B is the example for your Euclidean distance.

(Refer Slide Time: 11:21)

Mathematical Requirements of a Distance Function

- Both the Euclidean metric and the Manhattan metric satisfy the following mathematical requirements of a distance function, for all objects i, j, and h:
- (D1) $d(i, j) \geq 0$
- (D2) $d(i, i) = 0$
- (D3) $d(i, j) = d(j, i)$
- (D4) $d(i, j) \leq d(i, h) + d(h, j)$
- Condition (D1) merely states that distances are nonnegative numbers and (D2) says that the distance of an object to itself is zero
- Axiom (D3) is the symmetry of the distance function
- The triangle inequality (D4) looks a little bit more complicated, but is necessary to allow a geometrical interpretation
- It says essentially that going directly from i to j is shorter than making a detour over object h



The mathematical requirements of a distance function, both Euclidean metric and Manhattan metric, satisfy the following mathematical requirements of a distance function for all objects i, j and h. The first property is D1 $d(i, j) \geq 0$, $d(i, i) = 0$, $d(i, j) = d(j, i)$, $d(i, j) \leq d(i, h) + d(h, j)$. Condition D1 merely states that the distances are non-negative numbers and D2 says that the distance of an object itself is 0 because i, i is 0.

Axiom D3 is the symmetry of the distance function. The triangle inequality axiom D4 looks a little bit more complicated, but it is necessary to allow a geometrical interpretation. It says essentially that going directly from i to j is shorter than making a detour over object h. For example, suppose this is i this is j, and this is h so what says moving point i to j this will be shorter than moving i to h and h to j that is your triangular inequality.

(Refer Slide Time: 12:59)

Distances computation between the objects

- If $d(i, j) = 0$ does not necessarily imply that $i = j$, because it can very well happen that two different objects have the same measurements for the variables under study
- However, the triangle inequality implies that i and j will then have the same distance to any other object h , because $d(i, h) \leq d(i, j) + d(j, h) = d(j, h)$ and at the same time $d(j, h) \leq d(j, i) + d(i, h) = d(i, h)$, which together imply that $d(i, h) = d(j, h)$



15

Distance computation between the objects if d of $i, j = 0$ does not necessarily imply that $i = j$ because it can very well happen that two different objects have the same measurement for the variable under study. What is the meaning of this one is if the distance between object $i, j = 0$ it need not necessary that always it should be $i = j$. Sometimes there may be two objects which is not $i = j$ their distance also may be 0.

However, the triangle inequality implies that i and j will then have the same distance to any other object h because d of $i, h \leq d$ of $i, j + d$ of $j, h = d$ of j, h at the same time d of $j, h \leq d$ of $j, i + d$ of $i, h = d$ of i, h which together imply that d of $i, h = d$ of j, h .

(Refer Slide Time: 14:13)

Minkowski distance



- A generalization of both the Euclidean and the Manhattan metric is the Minkowski distance given by:

$$d(i, j) = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \dots + |x_{in} - x_{jn}|^p)^{1/p},$$

Where p is any real number larger than or equal to 1

- This is also called the L_p metric, with the Euclidean ($p = 2$) and the Manhattan ($p = 1$) as special cases



16

The next measure of the distance is Minkowski distance. A generalization of both Euclidean and Manhattan metric is the Minkowski distance. It is given by d of $i, j = \text{modulus of } (x_{i1} - x_{j1})^p + (x_{i2} - x_{j2})^p + \dots + (x_{in} - x_{jn})^p$ whole to the power $1/p$ where p is any real number larger than or equal to 1. This is also called the L_p metric for the Euclidean distance $p = 2$ and for Manhattan distance $p = 1$ as a special case.

(Refer Slide Time: 14:58)

Example for Calculation of Euclidean and Manhattan Distance

- Let $x_1 = (1, 2)$ and $x_2 = (3, 5)$ represent two objects as in the given Figure

The Euclidean distance between the two is $\sqrt{(2^2 + 3^2)} = 3.61$. The Manhattan distance between the two is $2 + 3 = 5$.

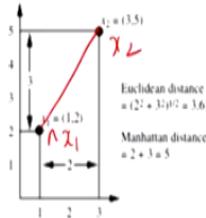


Figure: 4

Now let us take some example and calculate the Euclidean distance, Manhattan distance and Minkowski distance. Let $x_1 = (1, 2)$ and $x_2 = (3, 5)$ represent two objects this is point 1 so call it as x_1 this is point x_2 . The Euclidean distance between these two points x_1 x_2 is you see that it is $2^2 + 3^2$, square root it is 3.61. The Manhattan distance between the two points is this $2 + 3$. So this is your Euclidean distance, this is Manhattan distance.

You see that the Euclidean distance is smaller than the Manhattan distance because in Manhattan distance you cannot have a direct route, you have to take only a rectangular route that will be the larger. So this line represents Euclidean distance move here then move here when you add that, that represents your Manhattan distance.

(Refer Slide Time: 16:13)

n- by- n Matrix

- For example, when computing Euclidean distances between the objects of the following Table can be obtain as next slide:

- Euclidean distances between B and E:
 $((49 - 85)^2 + (156 - 178)^2)^{1/2} = 42.2$

Person	Weight(Kg)	Height(cm)
A	15	95
B	49	156
C	13	95
D	45	160
E	85	178
F	66	176
G	12	90
H	10	78



18

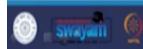
Let us take another example n- by-n matrix. This is one of the input for a cluster analysis for example, when computing Euclidean distance between the objects of the following table can be obtained in the next slides. For example, there are 1, 2, 3, 4, 5, 6, 7, 8. There are 8 persons their weight and heights are given. Now let us find out how to make n- by-n matrix, by calculating the distance between each persons each objects.

So generally, if you want to know the Euclidean distance between B and E, for example, B and E that is nothing but $(49 - 85)$ whole square. Otherwise, $(85 - 49)$ because we are squaring it + $(156 - 178)$ total square you take square root that is 42.2. So the distance between B and E is 42.2. Like that for between A and B, A and C we can find out.

(Refer Slide Time: 17:17)

n- by -n Matrix

	A	B	C	D	E	F	G	H
A	0	69.8	2.0	71.6	108.6	95.7	5.8	17.7
B	69.8	0	70.8	5.7	42.2	26.3	75.7	87.2
C	2.0	70.8	0	72.5	109.9	96.8	5.1	17.3
D	71.6	5.7	72.5	0	43.9	26.4	77.4	89.2
E	108.6	42.2	109.9	43.9	0	19.1	114.3	125.0
F	95.7	26.3	96.8	26.4	19.1	0	101.6	112.9
G	5.8	75.7	5.1	77.4	114.3	101.6	0	12.2
H	17.7	87.2	17.3	89.2	125.0	112.9	12.2	0



19

Do you see that n- by -n matrix the distance between in A and A is 0. You see that all the diagonal will be 0. The distance between A and B is 69.8 the distance between A and C is 2.0. So in my previous slides I have explained the distance between B and E is 42.2. So you see that this is symmetric see this upper triangle value is equal to your lower triangle value. So that is a replica, that is a mirror image of this value.

(Refer Slide Time: 17:54)

Interpretation

- The distance between object B and object E can be located at the intersection of the fifth row and the second column, yielding 42.2
- The same number can also be found at the intersection of the second row and the fifth column, because the distance between B and E is equal to the distance between E and B
- Therefore, a distance matrix is always symmetric
- Moreover, note that the entries on the main diagonal are always zero, because the distance of an object to itself has to be zero



20

Let us interpret this distance matrix. The distance between object B and E can be located at the intersection of the fifth row and the second column yielding 42.2. Now let us interpret that distance matrix. The distance between object B and E can be located at the intersection of fifth

row and second column. That was this one I am going to previous slide. The fifth row 1, 2, 3 ,4 fifth row second column this one 42.2.

The same number can be found at the intersection of 2nd row and 5th column because the distance between B and E is equal to the distance between E and B, therefore the distance matrix is always symmetric. Moreover, note that the entries of the main diagonal are always 0 because the distance of an object to itself has to be 0.

(Refer Slide Time: 18:54)

Distance matrix

- It would suffice to write down only the lower triangular half of the distance matrix

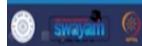
	A	B	C	D	E	F	G
B	69.8						
C	2.0	70.8					
D	71.6	5.7	72.5				
E	108.6	42.2	109.9	43.9			
F	95.7	26.3	96.8	26.4	19.1		
G	5.8	75.7	5.1	77.4	114.3	101.6	
H	17.7	87.2	17.3	89.2	125.0	112.9	12.2

Now we have shown only the lower triangle it would be suffice to write down only the lower triangular half of the distance Matrix.

(Refer Slide Time: 19:04)

Selection of variables

- It should be noted that a variable not containing any relevant information (say, the telephone number of each person) is worse than useless, because it will make the clustering less apparent.
- The Occurrence of several such “trash variables” will kill the whole clustering because they yield a lot of random terms in the distances, thereby hiding the useful information provided by the other variables.
- Therefore, such non informative variables must be given a zero weight in the analysis, which amounts to deleting them



22

Now let us see the selection of the variables, because before doing cluster analysis we have to see whether we have to select all the variables or the variables, which is relevant to our problem. It should be noted that a variable not containing any relevant information say the telephone number of each person is worse than useless because it will make the clustering less apparent. The occurrence of several such trash variable will kill the whole clustering.

Because they yield a lot of random terms in the distances thereby hiding the useful information provided by the other variables. Therefore, such non-informative variables must be given you zero weight in the analysis, which amounts to deleting them. So any not important variable, you can give zero weightage so that that will not be taken into calculation.

(Refer Slide Time: 20:02)

Selection of variables

- The selection of “good” variables is a nontrivial task and may involve quite some trial and error (in addition to subject-matter knowledge and common sense)
- In this respect, cluster analysis may be considered an exploratory technique



23

So the selection of good variable is a non-trivial task and may involve quite some trial and error in addition to subject matter knowledge and common sense. In this respect so a cluster analysis may be considered as an exploratory technique. In this lecture we have seen the effect of standardization then calculation of different types of distances with the help of example. I have explained how to find out Euclidean distance, Manhattan Distance and Minkowski distance.

Then formulation and interpretation of n by n matrix. Then I have explained this is one of the input for cluster analysis there are n objects, n variables, how to find out the distance between these two variables or objects. Then I have explained how to select relevant variables for the cluster analysis.

Lecture – 51
Clustering analysis: Part III

(Refer Slide Time: 00:46)

The slide has a blue header bar at the top and a blue footer bar at the bottom. The title 'Agenda' is centered in the header. The main content area contains a bulleted list of two items: 'Handling missing data' and 'Calculation of similarity and dissimilarity matrix'. In the footer bar, there are three small icons: a circular one, the word 'Swayam', and another circular one. On the right side of the footer bar, the number '2' is visible.

Agenda

- Handling missing data
- Calculation of similarity and dissimilarity matrix

In our previous class we have seen effect of standardization and how to find out different distances like Manhattan distance, Euclidean distance and Minkowski distance. Then I have explained how to select the variables. In this lecture we are going to see when you are collecting the data if you are missing some data, some data is not available how to handle that situation. Then very important concept of similarity and dissimilarity matrix. That is our agenda for this lecture.

(Refer Slide Time: 01:00)

Handling missing data

- It often happens that not all measurements are actually available, so there are some "holes" in the data matrix
- Such an absent measurement is called a missing value and it may have several causes
- The value of the measurement may have been lost or it may not have been recorded at all by oversight or lack of time



3

First let us see how to handle the missing data. It is often happens that not all measurements are actually available. So there are some holes in the data matrix that is a missing value in the data matrix. Such an absent measurement is called missing value it may have several causes. The value of measurement may have been lost or it may not have been recorded at all by oversight or lack of time.

Sometime the information is simply not available. For example, birth date of a foundling or the patients may not remember whether he or she ever had their measles, or it may be impossible to measure the desired quantity due to the malfunctioning of some instrument. In certain instances, the question does not apply or there may be more than one possible answer when the experiments obtain very different results.

(Refer Slide Time: 02:07)

Handling missing data

- How can we handle a data set with missing values?
- In a matrix we indicate the absent measurements by means of some code
- If there exists an object in the data set for which all measurements are missing, there is really no information on this object so it has to be deleted
- Analogously, a variable consisting exclusively of missing values has to be removed too

So how can we handle a data set with the missing values? That is important question now in a matrix we indicate that the absent measurement by means of some code. If there exists an object in the dataset for which all measurements are missing, there is really no information on this object, so it has to be deleted. Analogously a variable consisting exclusively of missing values has to be removed too.

(Refer Slide Time: 02:37)

Handling missing data

- If the data are standardized, the mean value m_f of the f^{th} variable is calculated by making use of the present values only
- The same goes for s_f ,

$$s_f = \frac{1}{n} \{ |x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f| \}$$

In the denominator , we must replace 'n' by the number of non missing values for that variable

- But of course only when the corresponding x_{if} is not missing itself

If the data are standardized, the mean value m_f of the f^{th} variable is calculated by making use of present values only. The same goes for your mean absolute deviation, so mean absolute deviation that is $S_f = (1 / n) \text{ modulus of } (x_{1f} - m_f) \text{ and so on } l x_{nf} - m_f l$. In the denominator, we must

replace n by the number of non-missing values for that variables, but of course only when the corresponding x_i is not missing itself.

(Refer Slide Time: 03:20)

Handling missing data

- In the computation of distances (based on either the x_i , or the z_i) similar precautions must be taken
- When calculating the distances $d(i, j)$, only those variables are considered in the sum for which the measurements for both objects are present subsequently the sum is multiplied by p and divided by the actual number of terms (in the case of Euclidean distances this is done before taking the square root)
- Such a procedure only makes sense when the variables are thought of as having the same weight (for instance, this can be done after standardization)

7

In the computation of distances based on the either X_i or the Z_i similar precautions must be taken when calculating the distances d of i, j only those variables are considered in the sum of which the measurements of both objects are present. Subsequently, the sum is multiplied by p and divided by the actual number of terms. In the case of Euclidean distances, this is done before taking the square root. Such a procedure only make sense when the variables are thought of as having the same weight. For instance, this can be done after standardization.

(Refer Slide Time: 04:03)

Handling missing data

- When computing these distances, one might come across a pair of objects that do not have any common measured variables, so their distance cannot be computed by means of the above mentioned approach.
- Several remedies are possible: One could remove either object or one could fill in some average distance value based on the rest of the data
- Or by replacing all missing x_{ij} by the mean mf of that variable; then all distances can be computed
- Applying any of these methods, one finally possesses a “full” set of distances

8

When computing these distances, one might come across a pair of objects that do not have any common measured variables, so their distance cannot be computed by means of above-mentioned approach. Several remedies are possible: One could remove either object, or one could feel some average distance value based on the rest of the data. Or, by replacing all missing x_{if} by the mean of m_f , that variable, then all distances can be computed. Applying any of these methods one finally possesses a full set of distances.

(Refer Slide Time: 04:44)

Dissimilarities

- The entries of a $n \times n$ matrix may be Euclidean or Manhattan distances
- However, there are many other possibilities, so we no longer speak of distances but of dissimilarities (or dissimilarity coefficients)
- Basically, dissimilarities are non-negative numbers $d(i, j)$ that are small (close to zero) when i and j are “near” to each other and that become large when i and j are very different
- We shall usually assume that dissimilarities are symmetric and that the dissimilarity of an object to itself is zero, but in general the triangle inequality does not hold



9

Then we will go to another topic that is dissimilarities. The entries of $n \times n$ matrix may be Euclidean or Manhattan distances. However, there are many other possibilities, so we no longer speak of distances, but dissimilarities or dissimilarity coefficients. Basically, dissimilarities are non-negative numbers that is d of i, j that are small, close to 0 when i and j are near to each other and they become large when i and j are very different. We shall usually assume that dissimilarities are symmetric and that the dissimilarity of an object to itself 0. But in general, the triangle inequality does not hold.

(Refer Slide Time: 05:40)

Dissimilarities

- Dissimilarities can be obtained in several ways.
- Often they can be computed from variables that are binary, nominal, ordinal, interval, or a combination of these
- Also, dissimilarities can be simple subjective ratings of how much certain objects differ from each other, from the point of view of one or more observers
- This kind of data is typical in the social sciences and in marketing



10

Dissimilarities can be obtained in several ways. Often, they can be computed from variables that are binary, nominal, ordinal, interval or combination of these. Also dissimilarities can be simple subjective rating of how much certain objects differ from each other from the point of view of one or more observers. This kind of data is typical in the social science or in the marketing.

(Refer Slide Time: 06:11)

Example

- Fourteen postgraduate economics students (coming from different parts of the world) were asked to indicate the subjective dissimilarities between 11 scientific disciplines.
- All of them had to fill in a matrix like Table 4, where the dissimilarities had to be given as integer numbers on a scale from 0 (**identical**) to 10 (very different)
- The actual entries of the Table in next slide, are the averages of the values given by the students



11

Let us take an example then I will explain the concept of dissimilarities fourteen post-graduate economic students coming from different parts of the world were asked to indicate the subjective dissimilarities between 11 scientific disciplines. All of them had to fill in a matrix, like in table 4 in the next slide where the dissimilarities had to be given as integer numbers, on a scale of 0 to

10, where the 0 represents identical 10 represents very different. The actual entries of the table in the next slides are the average of these values given by the students.

(Refer Slide Time: 06:54)

Example

- It appears that the smallest dissimilarity is perceived between mathematics and computer science (1.43), whereas the most remote fields were psychology and astronomy (9.36)

Astronomy	0.00
Biology	7.86 0.00
Chemistry	6.50 2.93 0.00
Computer sci.	5.00 6.86 6.50 0.00
Economics	8.00 8.14 8.21 4.79 0.00
Geography	4.29 7.00 7.64 7.71 5.93 0.00
History	8.07 8.14 8.71 8.57 5.86 3.86 0.00
Mathematics	3.64 7.14 4.43 1.43 3.57 7.07 9.07 0.00
Medicine	8.21 2.50 2.93 6.36 8.43 7.86 8.43 6.29 0.00
Physics	2.71 5.21 4.57 4.21 8.36 7.29 8.64 2.21 5.07 0.00
Psychology	9.36 5.57 7.29 7.21 6.86 8.29 7.64 8.71 3.79 8.64 0.00

12

It appears that the smallest dissimilarity is perceived between mathematics and computer science that value is 1.43 mathematics and computer science. This is our smallest dissimilarity whereas the most remote fields where psychology and astronomy psychology, astronomy. So this table represents dissimilarity matrix from that we can directly read, which is having lesser dissimilarity, which is having more dissimilarity.

(Refer Slide Time: 07:35)

Dissimilarities

- If one wants to perform a cluster analysis on a set of variables that have been observed in some population, there are other measures of dissimilarity
- For instance, one can compute the (parametric) Pearson product-moment between the variables f and g, or alternatively the (non-parametric) Spearman correlation

13

If one wants to perform a cluster analysis on a set of variables that have been observed in some population. There are other measures of dissimilarity. For instance, one can compute the Parametric Pearson product-moment between the variables f and g or alternatively non-parametric spearman correlation. Here the dissimilarity can be found with the help of your Pearson correlation or spearmen correlation. We know that the Pearson correlation is a parametric method spearmen correlation is non-parametric method. Because Spearman correlation is applicable only for ordinal data.

(Refer Slide Time: 08:18)

Dissimilarities

- Both coefficients lie between - 1 and + 1 and do not depend on the choice of measurement units
- The main difference between them is that the Pearson coefficient looks for a linear relation between the variables f and g, whereas the Spearman coefficient searches for a monotone relation

$$R(f, g) = \frac{\sum_{i=1}^n (x_{if} - m_f)(x_{ig} - m_g)}{\sqrt{\sum_{i=1}^n (x_{if} - m_f)^2} \sqrt{\sum_{i=1}^n (x_{ig} - m_g)^2}}$$

14

Both the coefficients lay between - 1 and + 1. Which one I am saying where our Pearson and Spearman correlation and do not depend on the choice of measurement units. We need not bother about the units because we are going to see the range of correlation coefficient is – 1 to + 1. Similarly, the Spearman correlation value also between -1 to + 1. That value does not depending upon what type of units of the data.

The main difference between is that the Pearson coefficients look for a linear relationship between variables f and g, whereas the spearmen coefficient searches for monotone relations. So, this is formula for our correlation coefficient, so we call it as r the correlation coefficient we have studied this formula already. So the correlation coefficient row is this is nothing but co variance, co variance of x, y divided by standard deviation of x and standard deviation of y. So this is in some other format this is x, y the correlation coefficient, not x, y here you can call it as f, g.

(Refer Slide Time: 09:40)

Dissimilarities

- Correlation coefficients are useful for clustering purposes because they measure the extent to which two variables are related
- Correlation coefficients, whether parametric or nonparametric, can be converted to dissimilarities $d(f, g)$, for instance by setting

$$d(f, g) = (1 - R(f, g))/2$$

With this formula, variables with a high positive correlation receive a dissimilarity coefficient close to zero, whereas variables with a strongly negative correlation will be considered very dissimilar



15

Correlation coefficients are useful for clustering purposes because they measures the extent to which two variables are related. Correlation coefficients, whether parametric or non-parametric can be converted into dissimilarities $d(f, g)$ for instance, by setting by this relationship. So dissimilarity between object $(f, g) = (1 - R$ that is correlation coefficient between $(f, g))$ divided by 2. iith this formula variables with a high positive correlation receive a dissimilarity coefficient close to zero whereas the variables with a strongly negative correlation will be considered as very dissimilar.

Why this kind of conversion is required the range of dissimilarity is between 0 to 1, but sometime what will happen the value of correlation coefficient between – 1 to + 1. So convert into to 0 to 1 scale we can use this transformation.

(Refer Slide Time: 10:46)

Similarities

- The more objects i and j are alike (or close), the larger $s(i, j)$ becomes
- Such a similarity $s(i, j)$ typically takes on values between 0 and 1, where 0 means that i and j are not similar at all and 1 reflects maximal similarity
- Values in between 0 and 1 indicate various degrees of resemblance
- Often it is assumed that the following conditions hold:

$$(S1) \ 0 \leq s(i, j) \leq 1$$

$$(S2) \ s(i, i) = 1$$

$$(S3) \ s(i, j) = s(j, i)$$

16

Now we enter into another concept called similarities. Previously we are explaining about dissimilarity and how we are going to study about what is similarities. The more objects i and j are alike, so the larger will be similarity between $s(i, j)$ becomes. Such a similarity s of (i, j) typically takes on values between 0 to 1 whereas 0 means that i and j are not similar at all, and 1 reflects maximum similarity. Values between 0 and 1 indicate various degrees of resemblance. Often it is assumed that the following conditions hold. So, S1: $0 \leq s(i, j) \leq 1$, because the range of similarities between 0 to 1. S2: the similarity between i, i itself 1, the similarity $s(i, j) = s(j, i)$

(Refer Slide Time: 11:48)

Similarities

- For all objects i and j , the numbers $s(i, j)$ can be arranged in an n -by- n matrix, which is then called a similarity matrix
- Both similarity and dissimilarity matrices are generally referred to as proximity matrices, or sometimes as resemblance
- In order to define similarities between variables, we can again resort to the Pearson or the Spearman correlation coefficient
- However, neither correlation measure can be used directly as a similarity coefficient because they also take on negative values

17

We will continue the concept of similarities for all objects i and j the numbers s of (i, j) can be arranged in an n -by- n matrix which is then called similarity matrix. Both similarity and dissimilarity matrices are generally referred to as proximity matrices sometimes as a resemblance. In order to define similarities between variables, we can again resort to a Pearson or Spearman correlation coefficient.

However, neither correlation measures can be used directly as a similarity coefficient because they also take on negative values because we cannot take the value of correlation and Spearman correlation as it is because they may range between - 1 to + 1, but the similarity values between 0 to 1.

(Refer Slide Time: 12:41)

Similarities

- Some transformation is in order to bring the coefficients into the zero-one range
- There are essentially two ways to do this, depending on the meaning of the data and the purpose of the application
- If variables with a strong negative correlation are considered to be very different because they are oriented in the opposite direction (like mileage and weight of a set of cars), then it is best to take something like the following:

$$s(f, g) = (1 + R(f, g))/2$$

which yields $s(f, g) = 0$ whenever $R(f, g) = -1$.

18

So in that case, we have to go for some transformation, some transformation is in order to bring the coefficients into the zero-one range. There are essentially two ways to do this, depending on the meaning of the data and the purpose of the application. If the variables with a strong negative correlation are considered to be very different because they are oriented in the opposite direction, like mileage and weight of a set of cars, then it is best to take something like the following.

You have to follow this transformation s of $(f, g) = (1 + R$ of $(f, g))/2$ what will happen here? We have added some constant so that constant will nullify the negative effect which yields the

similarity between f and $g = 0$ whenever the correlation coefficient is -1 because -1 and $+1$ becomes 0 . So this take care that the similarity value comes between 0 to 1 .

(Refer Slide Time: 13:49)

Similarities

- There are situations in which variables with a strong negative correlation should be grouped, because they measure essentially the same thing
- For instance, this happens if one wants to reduce the number of variables in a regression data set by selecting one variable from each cluster
- In that case it is better to use a formula like

$$s(f, g) = |R(f, g)|$$

which yields $s(f, g) = 1$ when $R(f, g) = -1$

19

There are situations in which variables with a strong negative correlation should be grouped because they measure essentially the same thing. For instance, this happens if one wants to reduce the number of variables in a regression dataset by selecting one variable from each cluster. In that case, it is better to use formula like this. Similarity between $(f, g) =$ the modulus value of correlation coefficient between f and g , which yields that the similarity between $(f, g) = 1$ when the correlation coefficient is -1 . We have to take only the positive values.

(Refer Slide Time: 14:31)

Similarities

- Suppose the data consist of a similarity matrix but one wants to apply a clustering algorithm designed for dissimilarities
- Then it is necessary to transform the similarities into dissimilarities
- The larger the similarity $s(i, j)$ between i and j , the smaller their dissimilarity $d(i, j)$ should be
- Therefore, we need a decreasing transformation, such as

$$d(i, j) = 1 - s(i, j)$$

20

Suppose the data consist of similarity matrix, but one wants to apply a clustering algorithm designed for dissimilarities. Then it is necessary to transform the similarities into dissimilarities. The larger the similarity the similarity between $s(i, j)$, between i and j the smaller their dissimilarity d of (i, j) should be. Therefore, we need a decreasing transformation. This is a very important result. So what it says that if you want to know the dissimilarity between two objects i, j that is nothing but $1 - \text{similarity between } i, j$

(Refer Slide Time: 15:13)

Binary Variables

- A contingency table for binary variables.

		object j		sum
		1	0	
object i	1	q	r	$q+r$
	0	s	t	$s+t$
sum		$q+s$	$r+t$	p



21

Let us take a binary type variable for that let us find the similarity and dissimilarity. Suppose a contingency table for binary variable is given. There is an object i and object j you see that object i is 1 0 there are two possibility object j also 1 0. So the q represents where the object i also takes value 1 object j also takes value 1. This r represents, q is the number of values here r represents $i = 1, j = 0$ this s represents number of values where $i = 0, j = 1$, t represents both $i = 1, j = 0$.

The row sum is $q + r$ for when $i = 1$, when $i = 0$ the row sum is $s + t$ same thing the column sum 's', when $j = 1$ the column sum is $q + s$, when $j = 0$ the column sum is $r + t$. So the sum of q, r, s, t that is nothing but your value p .

(Refer Slide Time: 16:30)

Dissimilarity between two binary variables

- $q \rightarrow$ is the number of variables that equal 1 for both objects i and j ,
- $r \rightarrow$ is the number of variables that equal 1 for object i but that are 0 for object j ,
- $S \rightarrow$ is the number of variables that equal 0 for object i but equal 1 for object j , and
- $t \rightarrow$ is the number of variables that equal 0 for both objects i and j .
- The total number of variables is p , where $p = q+r+s+t$.

What is the meaning of this q , r , s , t ? q represents the number of variables that equal 1 for both objects i and j you see that q is the number of variables r is the number of variables that equal one for object i but that are 0 for object j . S represents number of variables that equals 0 for object i , but equal 1 for object j . So t represents the number of variables that equals 0 for both objects i and j . The total number of variables is p where $p = q + r + s + t$.

(Refer Slide Time: 17:11)

Symmetric Binary Dissimilarity

$$d(i, j) = \frac{r+s}{q+r+s+t}.$$

The dissimilarity between symmetric binary variable, so from the previous table. What is the meaning of Symmetric binary variable is example is gender suppose 0 Male 1 female. You can reverse the code also there would not be any problem on this. So that is example of Symmetric Binary variable. So for Symmetric binary variable, how to find out dissimilarity. So dissimilarity

between i, j is it is $r+s$ what is $r+s$ we will go this. So this one, the dissimilarity is this value $r+s$ divided by sum of the all values $r+s$ divided by $q+r+s+t$.

(Refer Slide Time: 18:02)

Asymmetric binary variable

- A binary variable is asymmetric if the outcomes of the states are not equally important, such as the *positive* and *negative* outcomes of a disease test.
- By convention, we shall code the most important outcome, which is usually the rarest one, by 1 (e.g., *HIV positive*) and the other by 0 (e.g., *HIV negative*).
- Given two asymmetric binary variables, the agreement of two 1s (a positive match) is then considered more significant than that of two 0s (a negative match).
- Therefore, such binary variables are often considered “monary” (as if having one state).

24

Then let us see what is the meaning of Asymmetric binary variable. A binary variable is Asymmetric. If the outcomes of the states are not equally important, such as positive and negative outcome of disease test. By convention, we shall code the most important outcome, which is usually the rarest one by 1. For example, HIV positive that is the rarest one we will code it as 1 and the other by 0 HIV negative.

Given two Asymmetric binary variable the agreement of two 1s that is a positive match is considered more significant than that of two 0s that is negative match. Therefore, such binary variable are often considered as monary as if having only one state because we need not bother about the zero state, because zero state is that non-presence of HIV, because we are more concerned about presence of HIV where the state of one is more important.

(Refer Slide Time: 19:07)

asymmetric binary dissimilarity

		object j	
		1	0
object i	1	q	r
	0	s	t
sum	$q+s$	$r+t$	p

$$d(i, j) = \frac{r+s}{q+r+s}.$$



25

Now let us see how to find out dissimilarity value between Asymmetric binary variable. The same table which I have given contingency table which I have given previous table I have given. So the dissimilarity between Asymmetric matrixes d of (i, j). So we are considered about only r and s in this in the denominator there would not be t because we are not considering 0. So only we are writing q+ r + s. If it is Symmetric binary dissimilarity, the difference is there was a t element was there here. But here in the asymmetric binary dissimilarity formula there is no 't' element.

(Refer Slide Time: 19:51)

Jaccard coefficient

$$sim(i, j) = \frac{q}{q+r+s} = 1 - d(i, j).$$



26

Even we have seen this relationship, that relationship called the Jaccard co-efficient. That is the similarity between(i, j) = 1 - dissimilarity. So similarity q divided by (q + r+ s), where is q this

one, we are bothered about only the pretense of 1 so q divided by $q + r + s$ that is your similarity between i, j for a asymmetric binary variable. So if we want to know dissimilarity, that is simply the similarity equal to 1 minus dissimilarity between i and j .

(Refer Slide Time: 20:32)

Dissimilarity between binary variables								
name	gender	fever	cough	test-1	test-2	test-3	test-4	
Jack	M	Y	N	P	N	N	N	
Mary	F	Y	N	P	N	P	N	
Jim	M	Y	Y	N	N	N	N	
:	:	:	:	:	:	:	:	



27

Now let us take an example we will find out for Asymmetric binary variable how to find out dissimilarity matrix. This table shows there are different name is there Jack, Mary, Jim. There is a gender here gender is Symmetric binary variable. We are not going to consider this one because this is a different test fever, cough test 1, test 2, test 3, test 4. This is Asymmetric variable because where the presence of 1 is more important Y represents and P represents 1, N represents 0.

(Refer Slide Time: 21:19)

Dissimilarity between Jack and Marry

	name	gender	fever	cough	test-1	test-2	test-3	test-4
Jack	Jack	M	Y	N	P	N	N	N
Marry	Mary	F	Y	N	P	N	P	N
Jim	Jim	M	Y	Y	N	N	N	N
		1	1	1	1	1	1	1

Marry	1	0
1	2	1
0	0	3

$$d(Jack, Mary) = \frac{0+1}{2+0+1} = 0.33$$

28

For this matrix, let us find the dissimilarity matrix between Jack and Mary. So I brought the table again we will let us find out the dissimilarity matrices between Jack and Mary. So for Mary, there are two possibility 1 0 for Jack that is under 2 possibility, 1 and 0. So let us count how we got this 2 so Mary also 1 Jack also 1 there are two possibilities there Mary this is 1 possibility this is another possibility. So there is a two count, so we have written it as 2.

Then how we got this 1? Mary is 1 jack is 0 so that means this one where Mary is 1 Jack is 0. Now we will go this column where Mary is 0 Jack is 1 so Mary 0 is this one, I think there is no value for this. Let us see the last option that is Mary also 0 Jack is also 0, So this 1 2 this is 3 that is 3. So if you want to know the dissimilarity distance between Jack and Mary, so we know that the formula is so we will add this 0 + 1 divided by 2 + 0 + 1 so we got 0.33.

(Refer Slide Time: 22:50)

Dissimilarity between Jack and Jim								
	name	gender	fever	cough	test-1	test-2	test-3	test-4
Jim	Jack	M	Y	N	P	N	N	N
	Mary	F	Y	N	P	N	P	N
	Jim	M	Y	Y	N	N	N	N
	:	:	:	:	:	:	:	:
		1	0					
Jack	1	1	1					
	0	1	3					

$d(Jack, Jim) = \frac{1+1}{1+1+1} = 0.67$

29

Similarly, now let us find how to find out the dissimilarity between Jack and Jim. So Jack is taken in rows Jim is taken as in the column. So first we will find out how we got this 1. So this case is Jack also 1 Jim also 1. So this category so Jack also 1 Jim also 1. I think there is only one possibility, so it is 1 how we got this 1, the Jack is 1 Jim is 0. So this value Jack is 1 presence Jim is 0 that is 1. So how we got this 1 where Jack is 0 Jim is 1. So Jack is 0 yeah, this value Jack is 0 no means 0, Y means 1.

Let us see how we got this value 3 so Jack also is 0 Jim also is 0 so this 1 no this one 1, 2, 3 that is how we got the 3 values. So if you want to know the dissimilarity between Jack and Jim. So this is 1 + 1 divided by 1 + 1 + 1 so 2/3 it is 0.67.

(Refer Slide Time: 24:16)

Dissimilarity between Jim and Marry

		name	gender	fever	cough	test-1	test-2	test-3	test-4
		Jim							
		Marry							
		Jack	M	Y	N	P	N	N	N
		Mary	F	Y	N	P	N	P	N
		Jim	M	Y	Y	N	N	N	N
			:	:	:	:	:	:	:
				1	0				
		Marry	1	1	2				
		0		1	2				

$$d(Mary, Jim) = \frac{1+2}{1+1+2} = 0.75$$



30

Let us take another example the dissimilarity between Jim and Mary. So Mary there is two option 1 and 0. But Jim also there are two options 1 and 0 let us see how we got this value 1. Mary is 1, Jim also 1. So this possibility this one the second case Mary is 1, Jim is 0 this is one value. Mary is 1 Jim is 0, so there are two possibility. So that is where we got the value 2 how we got to this value 1, Mary is 0 Jim is 1, so Mary is 0 Jim = 1 that is this value.

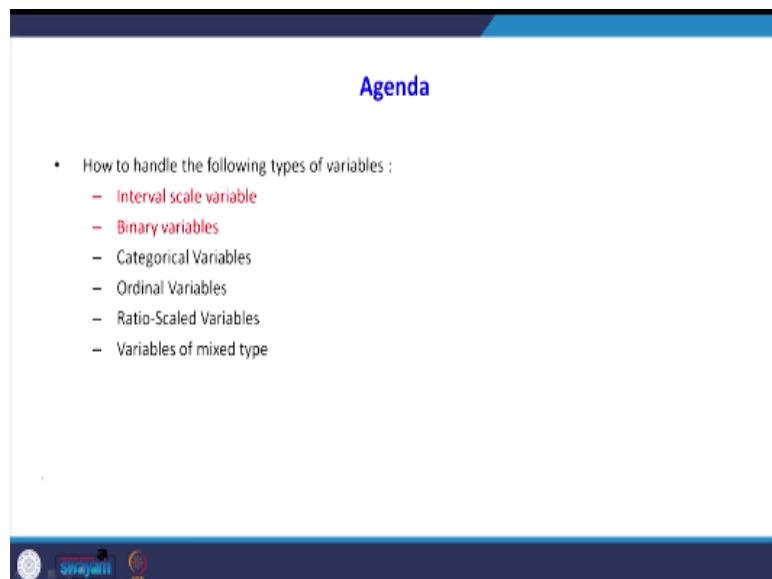
So how we got this 2 Mary is also 0 Jim also 0 so these two possibilities, Mary also 0 Jim also 0 test 1 here also Mary is 0 Jim also 0. So if you want to know asymmetric dissimilarity between Jim and Mary it is $1 + 2$ this value plus this one divided by this $1 + 1 + 2$ that is 4 so we got 0.75. So this is the way to find out asymmetric dissimilarity between different variables. In this class, we have seen how to handle the missing data for cluster analysis.

Then I have explained the concept of similarity and dissimilarity matrix. Then we have studied symmetric and asymmetric binary variable and how to find out the dissimilarity between symmetric binary variables and dissimilarity between asymmetric binary variables. Thank you.

Lecture – 52
Cluster Analysis - IV

In my previous lecture, I have explained how to handle missing data while doing clustering analysis. Second thing I have explained how to find dissimilarity and similarity matrix. The third one I have explained there is a binary variable, how to find out the dissimilarity and similarity matrix.

(Refer Slide Time: 00:46)



The screenshot shows a presentation slide with a blue header bar. The main title is 'Agenda' in blue text. Below it is a bulleted list of topics:

- How to handle the following types of variables :
 - Interval scale variable
 - Binary variables
 - Categorical Variables
 - Ordinal Variables
 - Ratio-Scaled Variables
 - Variables of mixed type



We will see how to handle other variables. For example, the agenda for these classes, if there is an interval scale variable how to use for that dataset for the cluster analysis and binary variable and how to use that dataset for the cluster analysis; we have done in our previous lecture. In this class, we will see; if there is a categorical variable how to use that dataset for clustering analysis. Next we will see there is ordinal variable, how to use for our cluster analysis.

And if there is Ratio-Scaled variables and how to do that data for our cluster analysis, and finally the data maybe mixed type, the combination of these above data, in that case how to use that dataset for our clustering analysis. The agenda for this lecture is how to handle the following

data types. One is interval-scaled variable and binary variable that I have covered in my previous lectures.

In this lecture if the variable nature is categorical, ordinal, ratio and combination of above dataset that is mixed type, let us see how to see this kind of variables and how to use these kind of variables for our clustering analysis.

(Refer Slide Time: 02:04)

The slide has a dark blue header and footer. The title 'Categorical Variables' is centered in the header in a light blue font. The main content area has a white background. It starts with a section titled 'Categorical Variables' in bold black font. Below it is a bulleted list:

- A categorical variable is a generalization of the binary variable in that it can take on more than two states
- For example, map color is a categorical variable that may have, say, five states: red, yellow, green, purple, and blue

Below the list is a grayscale map of the Western United States. Five states are highlighted with colored boxes: Nevada (yellow), Colorado (green), Wyoming (light green), Arizona (purple), and New Mexico (red). The footer contains three small icons: a person, a gear, and a question mark, followed by the number '3'.

First we will take an example, categorical variable. A categorical variable is a generalization of binary variable in that it can take on more than two states. It is similar to binary variable but in binary variable only 2 option will be there. But in categorical variable there maybe more than 2 states. For example, map color is a categorical variable that may have say 5 states red, yellow, green, purple and blue. This is an example of categorical dataset.

(Refer Slide Time: 02:36)

Categorical Variables

- Let the number of states of a categorical variable be M
- The states can be denoted by letters, symbols, or a set of integers, such as $1, 2, \dots, M$
- Notice that such integers are used just for data handling and do not represent any specific ordering

Let the number of states of a categorical variable be capital M . The states can be denoted by letters, symbols, or a set of integers such as $1, 2$ up to M . Notice that such integers are just used for data handling and do not represent any specific ordering.

(Refer Slide Time: 02:57)

Categorical Variables

- "How is dissimilarity computed between objects described by categorical variables?"

The question which we are going to answer in this lecture is how dissimilarity computed between objects described by categorical variables.

(Refer Slide Time: 03:05)

Categorical Variables

- The dissimilarity between two objects i and j can be computed based on the ratio of mismatches:

$$d(i, j) = \frac{p - m}{p},$$

where 'm' is the number of matches (i.e., the number of variables for which 'i' and 'j' are in the same state), and 'p' is the total number of variables

Weights can be assigned to increase the effect of 'm' or to assign greater weight to the matches in variables having a larger number of states

The dissimilarity between two objects i and j can be computed based on the ratio of mismatches. So the formula to find out the dissimilarity for a categorical variable is $d(i, j) = (p - m)$ divided by p, notice that it is a small m, m is the number of matches that is the number of variables for which i and j are in the same state and the p is the total number of variables which can be assigned to increase the effect of m or to assign greater weight to the matches in the variables having a larger number of states. So we can give weightage also for different states.

(Refer Slide Time: 03:48)

Dissimilarity between categorical variables

- Suppose that we have the sample data as shown in the table
- Let only the object-identifier and the variable (or attribute) test-1 are available, which is a categorical data

object identifier	test-1 (categorical)	test-2 (ordinal)	test-3 (ratio-scaled)
1	code-A	excellent	445
2	code-B	fair	22
3	code-C	good	164
4	code-A	excellent	1,210

Now we will take an example with the help of example I will explain how to find out the dissimilarity between categorical variables. So the source for this example is, this book finding groups in data and introduction to cluster analysis by Kaufman and Rousseeuw, the publisher is

John Wiley. Suppose that we have the sample data as shown in the table, the table having 1, 2, 3 there are 4 columns.

One is object identifier, second column is test-1 categorical data, third column is test-2 ordinal, the fourth one is test-3 ratio-scaled. Let only the object identifier and variable test-1 are available which is a categorical data, so for this example we are going to consider only 2 column, one is this object identifier, second one is the test-1 which is categorical data.

(Refer Slide Time: 04:47)

Dissimilarity matrix				
	1	2	3	4
1	0			
2		$d(2,1)$	0	
3		$d(3,1)$	$d(3,2)$	0
4		$d(4,1)$	$d(4,2)$	$d(4,3)$

The dissimilarity matrix, see that we are showing only the lower triangle that is 1, 2, 3, 4 is a object identifier, the diagonal will be 0 because the dissimilarity is 0 for this same value of when $i = j$. So this location is $d(2,1)$ second row first column, this location is third row first column third row second column, this location is fourth row first column fourth row second column and fourth row third column.

(Refer Slide Time: 5:18)

Dissimilarity between categorical variables

- Since here we have one categorical variable, test-1, we set $p = 1$ in Equation

$$d(i, j) = \frac{p - m}{p},$$

So that $d(i, j)$ evaluates to '0' if objects i and j match, and '1' if the objects differ

- Thus, we get

object Identifier (categorical)	test-1
1	code-A
2	code-B
3	code-C
4	code-A

9

Since here we have only one categorical variable, test-1, we set $p = 1$ in equation because p is the number of variables. So that $d(i, j)$ evaluate 0 if the object i and j match and 1 if the object differ, thus we get 0, 1 0, 1 1 0, 0 1 1 0. I will tell you how we got this matrix. Suppose let us find out how this location has come this value is 1, the distance between you see that this matrix $d(2, 1)$ so $d(2, 1)$ is we will use this $p - m$ divided by small p so p is number of variables because here only one variable minus m .

When you compare 1 and 2 because the 2, 1 or 2, 1 see code-A and code-B it is not matching so the value of m is 0 so $1 - 0$ divided by ($p = 1$), that is why we got this one value. Let us find out this value $d(4, 1)$. So there is a one variable is this 4 another variable this, so we will use the same formula $d(i, j)$ equal to; so $p = 1$, see code A and; for object identifier the code A is same for 1 and 4. So here the m is 1, so p also 1, m also 1, so this value is 0. This way the all other values were found.

(Refer Slide Time: 06:52)

Ordinal Variables

- A discrete ordinal variable resembles a categorical variable, except that the 'M' states of the ordinal value are ordered in a meaningful sequence
- Ordinal variables are very useful for registering subjective assessments of qualities that cannot be measured objectively *Very good, Good, bad*
- For example, professional ranks are often enumerated in a sequential order, such as Assistant, Associate, and full for Professors
- A continuous ordinal variable looks like a set of continuous data of an unknown scale; that is, the relative ordering of the values is essential but their actual magnitude is not

A handwritten note in red ink on a white background. It shows two pairs of numbers separated by a dash. The first pair is '99 - 1' and the second pair is '50 - 2'. Both pairs are enclosed in small rectangular boxes.

Now let us go to the next type of variable, ordinal variable. So ordinal variable is similar to categorical variable but the order is more important. So in my; when I am explaining the previous data that is a categorical variable one example is the pin code, for example in India the pin code is an example for our categorical data, so that number is not representing any meaning. So now we will start with ordinal variables.

A discrete ordinal variable resembles a categorical variable, except that M states of the ordinal values are ordered in a meaningful sequence, this term is very important because there is a ranking, there is an order in each value. Ordinal variables are very useful for registering subjective assessments of qualities that cannot be measured objectively. For example, say very good, good, bad so this way we can give the rank 1, 2, 3.

So here 1, 2, 3 says that is a rank for that. For example, professional ranks are often enumerated in a sequential order, such as Assistant, Associate, and full professors. So the order is more important. A continuous ordinal variable look like a set of continuous data of an unknown scale; that is the relative ordering of the values is essential but their actual magnitude is not. The problem with the ordinal scale in your class.

Suppose you are giving rank, those who got 99 is rank number 1, so those who are got number 2 50 marks rank number 2. See that this 1 and 2 signifies the rank but it is not the actual value

because this fellow got 99 marks, this fellow got 50 marks. So we are losing some important information when you go for ordinal dataset. Because for us this 1 and 2 is more important, how much mark they got is not important for us.

(Refer Slide Time: 08:58)

The slide has a blue header bar at the top and a blue footer bar at the bottom. The title 'Ordinal Variables' is centered in the header. The footer contains three small icons: a person, a book, and a gear, followed by the text 'Studydrive' and a copyright symbol. The main content area contains a bulleted list and a hand-drawn red step function diagram.

Ordinal Variables

- For example, the relative ranking in a particular sport (e.g., gold, silver, bronze) is often more essential than the actual values of a particular measure
- Ordinal variables may also be obtained from the discretization of interval-scaled quantities by splitting the value range into a finite number of classes
- The values of an ordinal variable can be mapped to ranks

A hand-drawn red step function diagram is shown, consisting of three horizontal steps of increasing height. Above the first step is the number '1', above the second is '2', and above the third is '3'. The steps are drawn with thick red lines on a white background.

For example, the relative ranking is a particular sport for example, gold, silver, bronze is often more essential than the actual value of a particular measures, because see there may be a different 3 scales, so rank 1, rank 2, rank 3 so this person gold, this fellow silver, this fellow bronze. Here what is more important is the; the rank is more important not the actual measures. So ordinal variable may also be obtained from discretization of intervals quantities by splitting the value range into finite number of classes.

Sometimes what happen, if there is a interval-scaled dataset that can be converted into ordinal variables. The values of ordinal variables can be mapped into ranks, so after converting ordinal then we can bring different ranks.

(Refer Slide Time: 09:52)

Dissimilarity computation

- The treatment of ordinal variables is quite similar to that of interval-scaled variables when computing the dissimilarity between objects
- Suppose that 'f' is a variable from a set of ordinal variables describing 'n' objects
- The dissimilarity computation with respect to 'f' involves the following steps:
- The value of 'f' for the i^{th} object is x_{if} , and 'f' has M_f ordered states, representing the ranking $1, \dots, M_f$.
- Replace each x_{if} by its corresponding rank, $r_{if} \in \{1, \dots, M_f\}$.

12

Let us see how to find out dissimilarity matrix for our ordinal dataset. The treatment of ordinal variable is quiet similar to that of interval-scaled variables when computing the dissimilarity between objects. Suppose that f is variable from a set of ordinal variables describing n objects. The dissimilarity computation with respect to f involves the following steps. The first one is the value of f for the ith object is $x(i, f)$ and f has M of ordered states, representing the ranking 1 to M.

So the M is the maximum number of rank. The $x(i, f)$ is the particular variable. So what we have to do we have to replace each $x(i, f)$ by its corresponding rank r if, so r if is the current rank the M_f is the maximum rank.

(Refer Slide Time: 10:51)

Dissimilarity computation

	A	B	C	D	E	F	G
B	69.8						
C	2.0	70.8					
D	71.6	3.5.7	72.5				
E	108.6	42.2	109.9	43.9			
F	95.7	26.3	96.8	26.4	19.1		
G	5.8	75.7	2.5.1	77.4	114.3	101.6	
H	17.7	87.2	17.3	89.2	125.0	112.9	12.2

13

Look at this data. This is; this also we have seen our previous lecture, this is Euclidean distance. So this is an example of dissimilarity computation. We have seen this previous table that is the Euclidean distances. This is an example of interval-scaled data. From this interval-scaled data we can convert this table into in ordinal dataset. So what we have to do, so suppose the lowest distance is highest rank.

So the lowest one is this one value, so this can be ranked as 1, the second one is 5., so this is rank 2. So next one is 5.7 rank 3, so; and so on, for each variable that is the interval dataset, you can convert into ordinal dataset by giving rank like 1, 2, 3 and so on. So the highest value will have the highest rank.

(Refer Slide Time: 11:52)

Standardization of ordinal variable

- Since each ordinal variable can have a different number of states, it is often necessary to map the range of each variable onto [0.0,1.0] so that each variable has equal weight.
- This can be achieved by replacing the rank r_{if} of the i^{th} object in the f^{th} variable by:

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

14

Standardization of ordinal variable. Since each ordinal variable can have a different number of states, it is often necessary to map the range of each variable onto 0 to 1 scale, so that each variable has equal weight. This can be achieved by replacing the rank r_{if} of the i^{th} object in the f^{th} variable by z_{if} equal to $(r_{if} - 1)$ divided by $(M_f - 1)$. So the r_{if} represents the current rank M_f represents the maximum rank.

(Refer Slide Time: 12:31)

Dissimilarity computation

- Dissimilarity can then be computed using any of the distance measures described earlier (like that for interval data)

15

Now let us see how to find out the dissimilarity computation. The dissimilarity can be computed using any of the distance measures described by earlier like that interval data.

(Refer Slide Time: 12:44)

Example

- Suppose that we have the sample data of the following Table ,
- Except that this time only the object-identifier and the continuous ordinal variable, test-2, are available
- There are three states for test-2, namely fair, good, and excellent, that is $M_f = 3$

object identifier	test-1 (categorical)	test-2 (ordinal)	test-3 (ratio-scaled)
1	code-A	excellent	445
2	code-B	fair	22
3	code-C	good	164
4	code-A	excellent	1,210



16

Now, let us take ordinal data the I will explain how to find out the dissimilarity matrix. Suppose that we have the sample data of following table, the same table which I have seen. So there are columns object identifier, test-1 categorical, test-2. Now we are going to consider these two column, one is object identifier next one is test-2 that is ordinal dataset. Suppose that we have the sample dataset for the following table except that is this time only the object identifier and the continuous ordinal variable, test-2 are available for us. There are 3 states of the test-2 namely good, excellent and fair, so the aim of three, because that is a maximum number of states.

(Refer Slide Time: 13:41)

Example

- For step 1, if we replace each value for test-2 by its rank, the four objects are assigned the ranks 3, 1, 2, and 3, respectively
- $$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$
- Step 2 normalizes the ranking by mapping rank 1 to 0.0, rank 2 to 0.5, and rank 3 to 1.0
- For step 3, we can use, say, the Euclidean distance, which results in the following dissimilarity matrix:



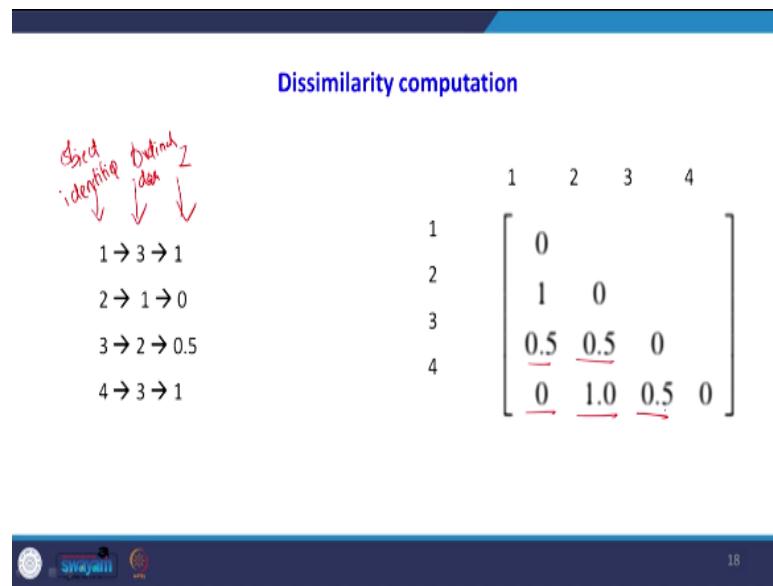
17

The step 1, if you replace each value of the test-2 by its rank, the four objects are assigned and the rank 3, 1, 2, 3 respectively. How is 3, 1, 2, 3? So this we are going to call test-3, this is also 3,

this is 1, this is 2. So how you are ranking this variable, 1 is fair, 2 is good, 3 is excellent. That is why we got this one, 3, 1; 3, 1, 2 and 3. Step 2, normalize the ranking by mapping 1 to 0, rank 2 to 0.5, rank 3 to 1. How we got this well, rank 1 = 0.

Because the r if value is 1 so $1 - 1$ divided by ; there are; the value of Mf is 3 so $3 - 1 = 2$; it is 0. The second one rank 2, $2 - 1$ divided by $3 - 1 = 2$ so what will happen, it is 1 divided by 2 it is a 0.5. The third one, so value of r if will be $3 - 1$ divided by this also $3 - 1$ so it is $2/2$ that is equal to 1. That is a way to standardize. Step 3, we can use say the Euclidean distance, which results the following dissimilarity matrix.

(Refer Slide Time: 15:16)



So this is our Euclidean distance. So the first column for example it says if the object identifier. This was our ranking ordinal dataset. This is our standardized value, so 3 is mapped into 2, 1; how we got this one see. $3 = 1$ so $1 = 0$, $2 = 0.5$, $3 = 1$. This is our standardized data. This is in the Z scale. This is our object identifier. This was our ordinal data. Now let us see how this matrix has come. Suppose if you want to know the distance between object identifier 2 and 1 so between 2 and 1 the distance is 1, so root of 1 square is 1.

So let us see how to find out distance between 3 and 1. So 3 is 1, so the formula is $1 - 0.5$ that is a 0.5 whole square. When you take square root this is 0.5. Let us see 3 to 2. So 3 to 2 $0 - 0.5$ whole square, square root that value is a 0.5. Let us see 4 to 1, what is the distance. So 4 to 1 is 1,

this distance is 1, $1 - 1$ that whole square, you take square root that as a 0. Let us see how we got this 1. So that is a 4 to 2. So 4 to 2 is 1 square, then square root that value is 1. So 4 to 3, so this distance is $0.5 - 1$ whole square, square root that value is 0.5. So this is an Euclidean distance. This is our standardized z value.

(Refer Slide Time: 17:18)

The slide has a dark blue header and footer. The title 'Ratio-Scaled Variables' is centered in a blue box. The main content area contains a bulleted list and a mathematical formula. Below the formula is a note about common examples. The footer contains icons for navigation and a page number '19'.

Ratio-Scaled Variables

- A ratio-scaled variable makes a positive measurement on a nonlinear scale, such as an exponential scale, approximately following the formula

$$Ae^{Bt} \quad \text{or} \quad Ae^{-Bt}$$

where A and B are positive constants, and t typically represents time

- Common examples include the growth of a bacteria population or the decay of a radioactive element

Now let us go to the next type of variable that is a Ratio-Scaled variable. A ratio-scaled variable makes a positive measurement on a nonlinear scale, such as an exponential scale, approximately following the formula. $A e^{Bt}$ or $A e^{-Bt}$ where, A and B are positive constants, and t is typically representing the time. So common example include the growth of a bacteria population or the decay of radioactive element, so that is an example of ratio-scale. Here the concept of ratio-scale is, there is a meaning for our absolute 0. When there is a absolute 0 you can do all kind of arithmetic operation using ratio-scaled data.

(Refer Slide Time: 18:08)

Computing the dissimilarity between objects

- There are three methods to handle ratio-scaled variables for computing the dissimilarity between objects:
 1. Treat ratio-scaled variables like interval-scaled variables
 - This, however, is not usually a good choice since it is likely that the scale may be distorted✓
 2. Apply logarithmic transformation to a ratio-scaled variable f having value x_{if} for object i by using the formula $y_{if} = \log(x_{if})$
 - The y_{if} values can be treated as interval valued. Notice that for some ratio-scaled variables, log or other transformations may be applied, depending on the variable's definition and the application

So computing the dissimilarity between the objects. There are three methods to handle the ratio-scaled variable for computing the dissimilarity between the objects. The first method is, treat ratio-scaled variable like interval-scaled variables. This, however, is not usually a good choice since it is likely that the scale may be distorted. But most of the marketing examples we will not differentiate the ratio-scale and interval-scale even though we collect the interval scale.

So we will use as a ratio-scale for finding all kind of statistical test. The second method is apply logarithmic transformation to a ratio-scaled variable f having the value x_{if} for a object i and using the formula $Y_{if} = \log(x_{if})$. It is nothing but if there is a ratio-scale just you take log of that, so that can be used for further analysis for finding the dissimilarity matrix. The Y_{if} values can be treated as interval valued, notice that for some ratio-scaled variables, so log of log or other transformation may be applied, depending upon the variables definition and the applications, because for we can use any kind of different transformations.

(Refer Slide Time: 19:31)

Computing the dissimilarity between objects

3. Treat x_{ij} as continuous ordinal data and treat their ranks as interval-valued

- The latter two methods are the most effective, although the choice of method used may depend on the given application



21

The third method is treat X if as a continuous ordinal data and treat their rank as interval-values. So this is the third method. The latter two methods are the most effective, although the choice of method may depend upon the given application.

(Refer Slide Time: 19:49)

Example

- This time, we have the sample data of the following Table,
- Except that only the object-identifier and the ratio-scaled variable, test-3, are available

object identifier	test-1 (categorical)	test-2 (ordinal)	test-3 (ratio-scaled)
1	code-A	excellent	445
2	code-B	fair	22
3	code-C	good	164
4	code-A	excellent	1,210



22

Now let us taken an example of ratio-scaled then I will explain how to find out the dissimilarity matrix. This time we have the sample data of the following table except the only object identifier, so we are going to consider this one and the ratio-scaled variable this column. So we are going to consider only these two column for finding the dissimilarity matrix, so in the third column ratio-scaled 445, 22, 164, 1210.

(Refer Slide Time: 20:23)

Example

- Let's try a logarithmic transformation 1 2 3 4
- Taking the log of test-3 results in the values 2.65, 1.34, 2.21, and 3.08 for the objects 1 to 4, respectively
- Using the Euclidean distance on the transformed values, we obtain the following dissimilarity matrix:

	1	2	3	4	
1	0				
2		0			
3			0		
4				0	

object identifier (ratio-scaled)	test-3
1	445
2	22
3	164
4	1,210



23

Let us try a logarithmic transformation. So just we are going to take the log of the column 3 values. Taking the log of test-3 results in the values become 2.65, 1.34, 2.21 and 3.08 for subject 1 to 4 respectively. When you look at this after taking log transformation the value it is compressed, see that it is scaled down; that is a purpose of scaling. So instead of using 445 you can use 2.65; it is easy to handle.

So instead of using 22 for cluster analysis application you can use 1.34. So the benefit of taking log of this one is it is compressed in a smaller scale. So using the Euclidean distance on the transformed values, we obtain the following dissimilarity matrix. So this is our dissimilarity matrix. So this is object identifier 1, 2, 3, 4. This is 1, 2, 3, 4. For example, $d(2, 1)$ how we got this one? This is 1, this is 2 so we have to find the difference.

We have to use this formula $(2.65 - 1.35)^2$ whole square, then square root. That value is 1.31. For example, 1 and 3 so this is 1 and 3 so the difference is $(2.65 - 2.21)^2$ whole square take square root so that will be this value. The suppose 4 1, so suppose this the fourth one, so find the difference $(3.08 - 2.65)^2$, then take square root so that value is 0.43. The same way we can get the other cells.

(Refer Slide Time: 22:05)

Variables of Mixed Types

- So far we have discussed how to compute the dissimilarity between objects described by variables of the same type, where these types may be either interval-scaled, symmetric binary, asymmetric binary, categorical, ordinal, or ratio-scaled
- However, in many real databases, objects are described by a mixture of variable types

Now we will enter into another type of variable; it is not another type of variable where whenever we do the cluster analysis there is a possibility of these variables like categorical, interval binary may come together. So that type of data types we are calling it is mixed types. So far we have discussed how to compute the dissimilarity between objects, described by variables of the same type, where these types may be either interval-scaled, symmetric binary, asymmetric binary, categorical, ordinal or ratio-scaled.

However, in reality, in many real databases, objects are described by the mixture of variable types. So whenever the mixture of these variable types are coming how to use the dataset, how to standardized that dataset for our further analysis of our cluster analysis, that we will see now.

(Refer Slide Time: 23:07)

Variables of Mixed Types

- In general, a database can contain all of the six variable types listed above
- "So, how can we compute the dissimilarity between objects of mixed variable types?"
- One approach is to group each kind of variable together, performing a separate cluster analysis for each variable type
 - This is feasible if these analyses derive compatible results
 - However, in real applications, it is unlikely that a separate cluster analysis per variable type will generate compatible results



25

In general, a database can contain all of 6 variable types listed above. So, how can we compute the dissimilarity between objects of mixed variable types? One approach is to group each kind of variables together, performing a separate cluster analysis for each variable type. This is feasible if this analysis derive compatible result. However, in real applications, it is unlikely that you separate cluster analysis per variable type will generate compatible result. So we can group the same variables, then you can go for cluster analysis. But sometime that will not be compatible. We cannot follow that approaches.

(Refer Slide Time: 23:52)

Variables of Mixed Types

- A more preferable approach is to process all variable types together, performing a single cluster analysis
- One such technique combines the different variables into a single dissimilarity matrix, bringing all of the meaningful variables onto a common scale of the interval [0.0,1.0]



26

A more preferable approach is to process all variable types together, performing a single cluster analysis. So in general what we have to do we have to by grouping all the variables we have to

do a single cluster analysis that will give you the meaningful result. One such technique combines the different variables into single dissimilarity matrix, bringing all of the meaningful variables onto a common scale of the interval 0 to 1. So one way to bring all different types of variables into a common scale is nothing but converting all the variables and bringing into this scale, nothing but standardization that is 0 to 1.

(Refer Slide Time: 24:36)

Variables of Mixed Types

- Suppose that the data set contains p variables of mixed type
- The dissimilarity $d(i, j)$ between objects i and j is defined as

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

where the indicator $\delta_{ij}^{(f)}$ = 0 if either

- x_{if} or x_{jf} is missing (i.e., there is no measurement of variable f for object i or object j), or $x_{if} = x_{jf} = 0$ and variable f is asymmetric binary;
- otherwise, $\delta_{ij}^{(f)} = 1$

27

Suppose that the data set contains p variables of mixed type, so the dissimilarity $d(i, j)$ between objects i and j is defined as $d(i, j) = (\sum_{f=1}^p d_{ij}^{(f)}) / (\sum_{f=1}^p \delta_{ij}^{(f)})$, for f where the indicator $d_{ij}^{(f)} = 0$, if either x_{if} or x_{jf} is missing that is there is no measurement of variables f for object i or object j or x_{if} equal to x_{jf} , equal to 0, then $d_{ij}^{(f)} = 0$ or the variable f is symmetric binary. Otherwise when the option is not there then the $d_{ij}^{(f)} = 1$.

(Refer Slide Time: 25:33)

Variables of Mixed Types

- The contribution of variable f to the dissimilarity between i and j, that is, $d_{ij}^{(f)}$, is computed dependent on its type:
- If 'f' is interval-based:
$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}},$$
 where h runs overall non missing objects for variable f
- If 'f' is binary or categorical: $d_{ij}^{(f)} = 0$, if $x_{if} = x_{jf}$
 - otherwise $d_{ij}^{(f)} = 1$

28

The contribution of variables f to the dissimilarity between i and j, that is, $d_{ij}^{(f)}$ is computed depending on its type. If 'f' is interval-based so $d_{ij}^{(f)} = \text{modulus of } (x_{if} - x_{jf}) \text{ divided by } (\max_h x_{hf} - \min_h x_{hf})$ where h runs overall non-missing objects for variable f. If f is binary or categorical so $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$ otherwise $d_{ij}^{(f)} = 1$.

(Refer Slide Time: 26:15)

Variables of Mixed Types

- If 'f' is ordinal: compute the ranks r_{if} and $z_{if} = r_{if} - 1 / M_f - 1$, and treat z_{if} as interval scaled
- If 'f' is ratio-scaled: either perform logarithmic transformation and treat the transformed data as interval-scaled; or treat 'f' as continuous ordinal data, compute r_{if} and z_{if} and then treat z_{if} as interval-scaled
- The above steps are identical to what we have already seen for each of the individual variable types

29

If f is ordinal compute the rank r if and z if using this formula, $z_{if} = (r_{if} - 1) / M_f - 1$, and treat z if as interval scaled. If f is ratio-scaled either perform logarithmic transformation and treat the transformed data as interval-scaled or treat f as continuous ordinal data, compute r if and z if and then treat z if as a interval-scaled. The above steps are identical to what we have already seen for each of the individual variable types.

(Refer Slide Time: 26:58)

Variables of Mixed Types

- The only difference is for interval-based variables, where here we normalize so that the values map to the interval [0,0,1,0]
- Thus, the dissimilarity between objects can be computed even when the variables describing the objects are of different types

The only difference is, for interval-based variables where here we normalize so that the values map to interval 0 to 1. Thus, the dissimilarity between objects can be computed even when the variables describe the objects are different types. The summary of this lecture is, I have explained how to handle the different types of variables. For example, if it is a categorical variable or ordinal variable and ratio-scaled variable and variables of mixed type, how to find out the dissimilarity matrix.

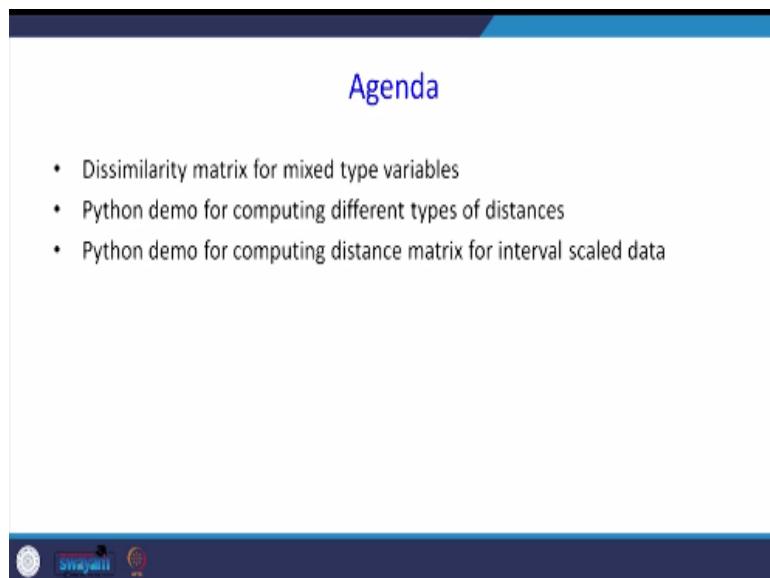
So what I have done in this lecture, I have taken one example in that using that example I have explained how to find out the dissimilarity matrix. But only for the last that variables of mixed type I have explained only the theory portions. The next class I will take another example which are mixed in nature then I will tell you how to find out the dissimilarity matrix. Along with that, we will start a new topic in the next lecture that is a K means algorithm.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 53
Cluster Analysis - V

In our previous class, I have explained how to handle different types of data for doing cluster analysis. I have started some theory, there is a mixed data type, how to handle that kind of data.

(Refer Slide Time: 00:38)



The slide has a dark blue header bar at the top and bottom. The main content area is white with a title 'Agenda' in blue. Below the title is a bulleted list of three items:

- Dissimilarity matrix for mixed type variables
- Python demo for computing different types of distances
- Python demo for computing distance matrix for interval scaled data

At the bottom of the slide, there is a dark blue footer bar with three small icons: a circular logo, the word 'swayam', and another circular logo.

The agenda for today's lecture is how to find the dissimilarity matrix for mixed type variables. And we will do python demo for computing different types of distances which I have explained theory in my previous lectures and also I will tell you python demo for computing distance matrix for interval scaled data.

(Refer Slide Time: 01:01)

Example

Consider the data given in the following table and compute a dissimilarity matrix for the objects of the table.
Now we will consider all of the variables, which are of different types

object identifier	test-1 (categorical)	test-2 (ordinal)	test-3 (ratio-scaled)
1	code-A	excellent	445
2	code-B	fair	22
3	code-C	good	164
4	code-A	excellent	1,210



3

Now let us take an example. This is a mixed type dataset. So consider the data given in the following table and compute a dissimilarity matrix for the objects of the table. Now we will consider all of the variables which are different types. So there are three type of dataset one is categorical, ordinal and ratio-scaled. If this kind of mixed data type is there how to use these kind of dataset for finding dissimilarity matrix and giving as an input for the cluster analysis.

(Refer Slide Time: 01:34)

Example

- The procedures we followed for test-1 (which is categorical) and test-2 (which is ordinal) are the same as outlined above for processing variables of mixed types
- For categorical variable - $d(i, j) = \frac{p - m}{p}$,
- For ordinal variable - $z_{ij} = \frac{r_{ij} - 1}{M_j - 1}$
- For interval scale variable - $d_{ij}^{(f)} = \frac{|x_{ij} - x_{jf}|}{\max_{k} x_{kj} - \min_{k} x_{kj}}$,



4

The procedures we followed for the test-1 which is categorical data and test-2 which is ordinal data are the same as outlined above for processing variables of mixed type. So what we have done, if it is a categorical variable type we have used this $p - m$ divided by p so where the p is number of variables where m is number of matches, that we have discussed in our previous class.

If it is an ordinal data, we have to standardize into 0 to 1 by using this formula $z_{if} = (r_{if} - \text{current rank}) / (\text{maximum rank} - 1)$. So this after converting into 0 to 1 scale then you can use our simple Euclidean method for finding the dissimilarity matrix. For interval scale variable so $d^{(f)}_{ij} = \text{modulus of } (x_{if} - x_{jf}) / (\max_h x_{hf} - \min_h x_{hf})$.

(Refer Slide Time: 02:33)

Normalizing the interval scale data

- First, however, we need to complete some work for test-3 (which is ratio-scaled)
- We have already applied a logarithmic transformation to its values
- Based on the transformed values of 2.65, 1.34, 2.21, and 3.08 obtained for the objects 1 to 4, respectively, we let $\max_h x_h = 3.08$ and $\min_h x_h = 1.34$
- We then normalize the values in the dissimilarity matrix obtained in [Example solve for ratio data](#) by dividing each one by $(3.08 - 1.34) = 1.74$

First normalizing the interval scale data. First, however, we need to complete some work for test-3 which is ratio-scaled. We have already applied a logarithmic transformation to its values. Based on the transformation values we got 2.65, 1.34, 2.21 and 3.08 obtained for the objects 1 to 4, respectively. We let maximum value of hx is from this among these values the maximum value is 3.08 and the minimum value is 1.34. The normalized value in the dissimilarity matrix obtained in the example and solve for ratio data by dividing each one by this difference that is $(3.08 - 1.34)$ that is 1.74.

(Refer Slide Time: 03:21)

Dissimilarity matrix for test-3

- This results in the following dissimilarity matrix for test-3:

Object Identifier	Ratio scaled Data (x)	Log (x)	
1	445	2.65	
2	22	1.34	
3	164	2.21	
4	1210	3.08	

→

$$\begin{bmatrix} 0 & & & \\ 0.75 & 0 & & \\ 0.25 & 0.50 & 0 & \\ 0.25 & 1.00 & 0.50 & 0 \end{bmatrix}$$

- For 1 and 2 = $(2.65 - 1.34) / (3.08 - 1.34) = 0.75$



6

This results in the following dissimilarity matrix for test-3. So what happened there was a object identifier was there, there was a ratio-scale was there, I am calling it as x. So we have taken log of that value 2.65, 1.34, 2.21. So we have to standardize this one for that purpose for example 1 and 2 we use this formula that is this formula to standardize. So here 2.65 this value minus this value divided by the maximum value and minus minimum value.

So when you divide this, this is a 0.75. So 2, 1 this was the distance. Similarly, if you want know 4, 1 so 4, 1 so the difference is $2.65 - 3.08$ whole divided by this value, that is 1.74 will give you this value. The same way the standardization is done for all the cells.

(Refer Slide Time: 04:23)

dissimilarity matrices for the three variables

- We can now use the dissimilarity matrices for the three variables in our computation of Equation $d_{ij}^{(f)} = \frac{|x_{ij} - x_{if}|}{\max_{k} |x_{kj} - \min_k |x_{kj}|}$,
- For example, we get $d(2,1) = (1(1) + 1(1) + 1(0.75)) / 3 = 0.92$

$$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

Dissimilarity matrix
for categorical

$$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0.5 \\ 0.5 \\ 0 \\ 0 \\ 1.0 \\ 0.5 \\ 0 \end{bmatrix}$$

Dissimilarity matrix
for ordinal

$$\begin{bmatrix} 0 \\ 0.75 \\ 0 \\ 0.25 \\ 0.50 \\ 0 \\ 0.25 \\ 1.00 \\ 0.50 \\ 0 \end{bmatrix}$$

normalize the values in the
dissimilarity matrix for ratio data



7

Now let us consider dissimilarity matrices for all three variables. What are the three variables? One is categorical, ordinal and ratio data. So we can now use the dissimilarity matrices for the three variables in our computation of equation, so $d^{(f)}_{ij} = \text{modulus of } (x_{if} - x_{jf}) \text{ divided by } (\max_h X_{hf} - \min_h X_{hf})$.

(Refer Slide Time: 04:51)

Example

- The resulting dissimilarity matrix obtained for the data described by the three variables of mixed types is:

$$(2,1) \begin{bmatrix} 0 & & & \\ 0.92 & 0 & & \\ 0.58 & 0.67 & 0 & \\ 0.08 & 1.00 & 0.67 & 0 \end{bmatrix}$$

For example, we got between $d_{2,1}$. So this is the location, 2, 1. I will explain how we got 0.92. So how we got this one is, so there are three variables is there; for dissimilarity matrix it is 1, for categorical variable, for ordinal variable, for 2, 1 portion is 1, for ratio data it is a 0.75. So we will find out the weighted mean. So weight is we are giving equal weigh for all kind of dataset. So $(1 \text{ into } 1) + (1 \text{ into } 1) + (1 \text{ into } 75)$ so total sum of weight is 3, so 0.92 that is why we got this value.

Similarly, each element you can do that one, for example here $1 \text{ into } 1 + 1 \text{ into } 0.5 + 1 \text{ into } 0.25$ divided by 3 so we will get this value. So this matrix is a resulting dissimilarity matrix attained for data described by the 3 variables for the mixed type. So this dissimilarity matrix is given as a input for doing the cluster analysis. So what happened this is a combined matrix, combined dissimilarity matrix for all three kind of variables, so what are that variables it is for categorical, ordinal and ratio data.

(Refer Slide Time: 06:12)

Interpretation

- If we go back and look at Table of given data, we can intuitively guess that objects 1 and 4 are the most similar, based on their values for test-1 and test-2
- This is confirmed by the dissimilarity matrix, where $d(4,1)$ is the lowest value for any pair of different objects
- Similarly, the matrix indicates that objects 2 and 4 are the least similar



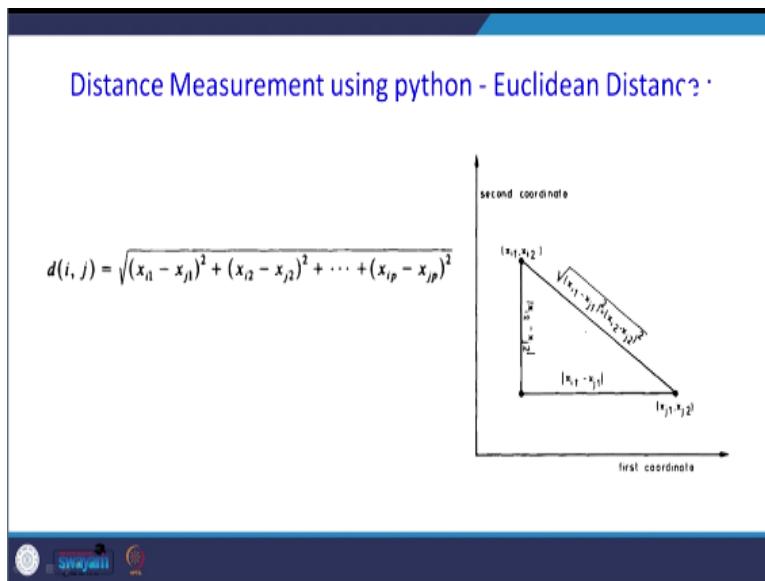
9

If you go back and look at the table of the given data we can intuitively guess that the object 1 and 4 are most similar, based on their values for test-1 and test-2. You see that, because if you look at this one, this is 0, this is 0, this is 0.25. In the given matrix, for example, for say for categorical data it is 0, for ordinal data it is 0, it look like the position of 4, 1 for both the variables is seems to be very similar. But what is happening here if the third where it is ratio data it is 0.25.

But when you find the average of 3 weighted average by looking at the dissimilarity matrix for categorical and ordinal look in the position of 4, 1. So this seems to be very similar because it is close to 0. But what is happening the 0.25 it is not very close to 0, so this can be verified by when you look at this one, so among the all the dataset in the combined the resulting dissimilarity matrix the value of 0.0 it is very similar, so the position of 4, 1 is very similar to each other, that is the point.

This is confirmed by the dissimilarity matrix, where 4, 1 is the lowest value for any pair of objects. Similarly, the matrix indicates that the object 2, 4 are the least similar. When you look at this 2, 4, so the highest value is 1. When you go there, here also 2, 4 here also it is 1, here also 1, here also 1 so in the resulting dissimilarity matrix also this seems to be 1 so highly dissimilar.

(Refer Slide Time: 07:55)



Now, we will go to the distance measurement using python. In my previous class, I have explained how to find out Euclidean distance, Manhattan distance, Minkowski distance. So this is an example for Euclidean distance, the formula for finding Euclidean distance is $(x_{i1} - x_{j1})$ whole square + $(x_{i2} - x_{j2})$ whole square + $(x_{ip} - x_{jp})$ whole square, then square root. If there are only two variable the distance formula is $(x_{i1} - x_{j1})$ whole square + $(x_{i2} - x_{j2})$ whole square, then square root.

(Refer Slide Time: 08:30)

Python Demo for Euclidean Distance

```

In [1]: import scipy
         from scipy.spatial import distance

#Euclidean Distance

In [2]: import numpy as np
         a = [1,2,3]
         b = [4,5,6]
         dst = distance.euclidean(a,b)

In [3]: dst
Out[3]: 5.196152422706632

```

But how to use python command, I brought the screenshot of that, for that import scipy, from scipy.spatial import distance, so how to find the Euclidean distance? So import numpy as np, so a and b there are two; a is one point where 1, 2, 3; the b is another point 4, 5, 6. If you want to

know the distance Euclidean distance between a, b so we have to write `dst = distance.euclidean(a, b)`. So when you type `dst`, so this is the our Euclidean distance between point a and b.

(Refer Slide Time: 09:07)

The slide has a blue header bar with the title 'Distance Measurement using python – Minkowski Distance :'. Below the title is a mathematical formula for the Minkowski distance:

$$d(i, j) = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{in} - x_{jn}|^p)^{1/p},$$

with a note below it:

- $p=1$ Manhattan distance
- $p=2$ Euclidean distance

At the bottom of the slide is a navigation bar with icons for back, forward, and search, and the text 'Swayam'.

The next, the distance is Minkowski distance. So the Minkowski distance is, the combination of both Manhattan distance and Euclidean distance. So $d(i, j) = |x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{in} - x_{jn}|^p$ whole to the power $1/p$. So what is happening, this is small p. If $p = 1$, then it is a Manhattan distance. If the $p = 2$ it is an Euclidean distance. Let us see how to find out this Minkowski distance using python.

(Refer Slide Time: 09:46)

The cell has a title 'Python Demo for Minkowski Distance'. It contains the following Python code and its output:

```
#Minkowski Distance
In [4]: distance.minkowski([1, 0, 0], [0, 1, 0], 1) #manhattan distance
Out[4]: 2.0

In [5]: distance.minkowski([1, 0, 0], [0, 1, 0], 2) #Euclidean distance
Out[5]: 1.4142135623730951

In [6]: distance.minkowski([1, 2, 3], [4, 5, 6], 2)
Out[6]: 5.196152422706632

In [7]: distance.minkowski([1, 2, 3], [4, 5, 6], 3)
Out[7]: 4.326748710922245
```

At the bottom of the cell is a navigation bar with icons for back, forward, and search, and the text 'Swayam'.

So I have taken two points, Minkowski distance let us see 1, 0, 0; 0, 1, 0. So the comma 1, this represent that we are finding Manhattan distance. So when you enter this the Manhattan distanced is 2. So this same dataset if you type 2, you will get Euclidean distance because the p = 2 will get you formula for the Euclidean distance that is 1.14. So this was our another example just to verifying this (1, 2, 3); (4, 5, 6), 2 so this will represent our Euclidean distance.

Suppose if you take 1, 2, 3 and 4, 5, 6 the p value can be 3 also; if it is 3 then Minkowski distance is 4.32. This formula distance.minkowski (1, 2, 3); (4, 5, 6), 2 this data previously we have used the distance between by using this formula distance.euclidean distance a, b we got 5.19. So by using Minkowski formula also the Minkowski formula the same dataset if you type 2 you will get the same answer. So the distance.minkowski between the two points (1, 2, 3) and (4, 5, 6), 3 where the p = 3 so this our Minkowski distance.

(Refer Slide Time: 11:07)

```

Dissimilarity matrix

In [9]: #Assimilarity or distance matrix
import pandas as pd
from scipy.spatial import distance_matrix

data = [[1, 3], [2, 5], [3, 6]]
df = pd.DataFrame(data, columns=['a', 'b'])
df

Out[9]:
   a   b
0  1  3
1  2  5
2  3  6

In [10]: pd.DataFrame(distance_matrix(df.values, df.values))
Out[10]:
   0   1   2
0  0.000000  1.414214  2.828427
1  1.414214  0.000000  1.414214
2  2.828427  1.414214  0.000000

```

Now, can we explain how to find out dissimilarity matrix? So for that you have to import pandas as pd from scipy.spatial import distance_matrix. So there are 1, 2; there are 3 points; 1, 4; 2, 5; 3, 6. So there are two columns a, b. So pd.DataFrame data, columns equal to this one, you will get this kind of output. So if you want to know the distance between a, b so there are identifier name is 0 1 and 2. There are two variables a and b. So if you want to know the distance matrix between different identifier 0 0 is 1; 1 0 is 1.41; 2 0 is 2.84 by using this command distance_matrix df.values and df.values.

(Refer Slide Time: 12:05)

Distance matrix calculation for Interval-Scaled Variables

- For example :
- Take eight people, the weight (in kilograms) and the height (in centimetres)
- In this situation, n = 8 and p = 2.

Person	Weight(Kg)	Height(cm)
A	15	95
B	49	156
C	13	95
D	45	160
E	85	178
F	66	176
G	12	90
H	10	78

Now, distance matrix calculation for interval-scaled variables. For example, there are 1, 2, 3, 4, 5, 6, 7, 8 there is a 8 persons, we can say object identifier, take 8 people, the weight is given in kilograms and the height is given in centimeters, so n = 8, so p = 8. Now how to find out the distance matrix?

(Refer Slide Time: 12:30)

```
#data matrix  
  
In [5]: import pandas as pd  
from scipy.spatial import distance_matrix  
  
data = [[15, 95], [49, 156], [13, 95], [45, 160], [85, 178], [66, 176], [12, 90], [10, 78]]  
ctys = ['A', 'B', 'C', 'D', 'E', 'F', 'G', 'H']  
df = pd.DataFrame(data, columns=['weight', 'Height'], index=ctys)
```

```
In [6]: df  
Out[6]:
```

	Weight	Height
A	15	95
B	49	156
C	13	95
D	45	160
E	85	178
F	66	176
G	12	90
H	10	78

So I have taken the data 15, 95; 49, 156 and so on. So there is a labels A, B, C, D, E, F, G, H. So the data frame is so Weight and Height. So we got this table. So weight and height is there, these are different identifier.

(Refer Slide Time: 12:54)

```
In [7]: Distance_matrix = pd.DataFrame(distance_matrix(df.values, df.values), index=df.index, columns=df.index)
```

```
Out[7]:
```

	A	B	C	D	E	F	G	H
A	0.000000	69.835521	2.000000	71.500105	106.577162	95.718337	5.830952	17.720045
B	69.835521	0.000000	70.830784	5.656854	42.190046	26.248809	75.663730	87.206651
C	2.000000	70.830784	0.000000	72.449983	109.077204	95.786760	5.099020	17.262677
D	71.500105	5.656854	72.449983	0.000000	43.863424	26.400758	77.386630	89.157165
E	106.577162	42.190046	109.077204	43.863424	0.000000	19.104973	114.337221	125.000000
F	95.718337	26.248809	95.786760	26.400758	19.104973	0.000000	101.548018	112.971608
G	5.830952	75.663730	5.099020	77.386630	114.337221	101.548018	0.000000	12.165525
H	17.720045	87.206651	17.262677	89.157165	125.000000	112.971608	12.165525	0.000000



Suppose if you want to know the distance matrix, so for that pd.DataFrame distance_matrix you take these values. So we are getting the distance matrix between A and A, 0, B and A. So you see that the diagonal value 0 because it is a Replica of because the distance between F and F is 0; G and G 0, H and H is 0. So what is happening between A and; B and A it is 69.83; A and B also 69.83 just a mirror value.

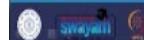
(Refer Slide Time: 13:29)

Distance matrix calculation using Python

```
In [8]: Distance_matrix.round(decimals=1, out=None)
```

```
Out[8]:
```

	A	B	C	D	E	F	G	H
A	0.0	69.8	2.0	71.6	108.6	95.7	5.8	17.7
B	69.8	0.0	70.8	5.7	42.2	26.2	75.7	87.2
C	2.0	70.8	0.0	72.4	109.9	96.8	5.1	17.3
D	71.6	5.7	72.4	0.0	43.9	26.4	77.4	89.2
E	108.6	42.2	109.9	43.9	0.0	19.1	114.3	125.0
F	95.7	26.2	96.8	26.4	19.1	0.0	101.5	112.9
G	5.8	75.7	5.1	77.4	114.3	101.5	0.0	12.2
H	17.7	87.2	17.3	89.2	125.0	112.9	12.2	0.0



We will round into 1 decimal. So by using matrix; distance_matrix.round (decimals = 1, out = none) so we got this matrix. This will run with the help of python.

(Refer Slide Time: 13:41)

```

jupyter Distance_measure (untrusted changes)

File Edit View Insert Cell Kernel Widgets Help
Trusted Python 3.0

In [1]: %import scipy
         from scipy.spatial import distance

#Euclidean Distance

In [2]: %import numpy as np
         a = [1,2,3]
         b = [4,5,6]
         dst = distance.euclidean(a,b)

In [3]: dst
Out[3]: 5.196152422706632

#Minkowski Distance

In [4]: distance.minkowski([1, 0, 0], [0, 1, 0], 1) #Manhattan distance

```

Now let us see how to use python to find the distance between points. So import spicy from scipy.spatial import distance, I will run that one. Now let us know how to find out Euclidean distance. So import numpy as np, there are two point a, b. The position of point a is 1, 2, 3; the position of point b is 4, 5, 6. If you want to know the distance so dst = distance.euclidean then if you want to know the what is the distance so will let us display this, so distance is 5.19 that is the between point a and b the Euclidean distance is 5.19.

(Refer Slide Time: 14:27)

```

jupyter Distance_measure (untrusted changes)

File Edit View Insert Cell Kernel Widgets Help
Trusted Python 3.0

In [1]: %import scipy
         from scipy.spatial import distance

In [2]: %import numpy as np
         a = [1,2,3]
         b = [4,5,6]
         dst = distance.euclidean(a,b)

In [3]: dst
Out[3]: 5.196152422706632

#Minkowski Distance

In [4]: distance.minkowski([1, 0, 0], [0, 1, 0], 1) #Manhattan distance
Out[4]: 2.0

In [5]: distance.minkowski([1, 0, 0], [0, 1, 0], 2) #Euclidean distance
Out[5]: 3.4142135623730951

In [6]: distance.minkowski([1, 2, 3], [4, 5, 6], 2)
Out[6]: 5.196152422706632

In [7]: distance.minkowski([1, 2, 3], [4, 5, 6], 3)

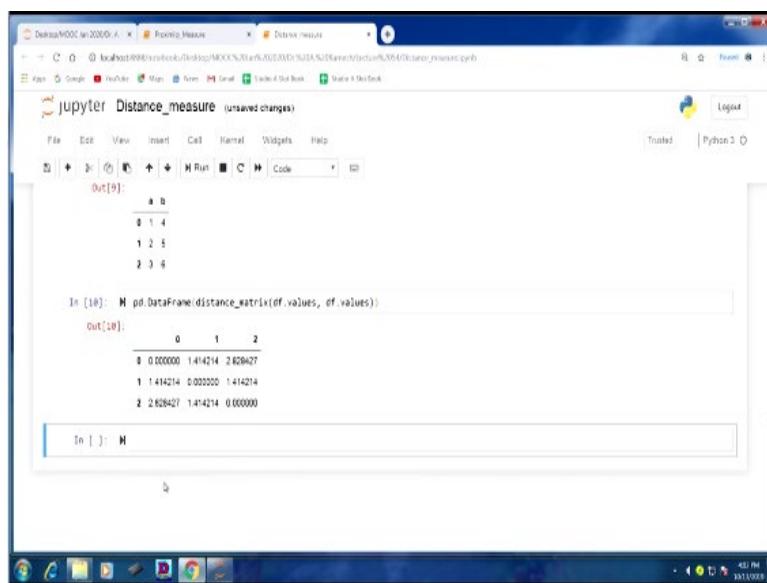
```

Now, let us know how to find out the Minkowski distance. For that, the Minkowski distance is calculated by using this function distance.minkowski. This is position of point A and position of point B. Then if you use 1 you will get a Manhattan distance. So this is Manhattan distance.

Instead of 1 if use 2 you will get here. Instead of 1 if you get 2 you will get a Euclidean distance. This is 1.41. We already got 5.19 as Euclidean distance even the Minkowski function you can get; you can verify that answer.

What happened here I gave distance.minkowski, I taken the same point that is 1, 2, 3 and 4, 5, 6 this one I used 2, so number 2 is used to get the Euclidean distance. Number 1 is use to get Manhattan distance. So we got to see that this also 5.19 when in the function Minkowski for use p, otherwise by using distance.euclidean function we got 5.19, so both are same.

(Refer Slide Time: 15:43)



The screenshot shows a Jupyter Notebook interface with two code cells. The first cell contains:

```
In [9]: a b  
0 1 4  
1 2 5  
2 3 6
```

The second cell contains:

```
In [10]: pd.DataFrame(distance_matrix(df.values, df.values))  
Out[10]:  
   0   1   2  
0 0.000000 1.414214 2.828427  
1 1.414214 0.000000 1.414214  
2 2.828427 1.414214 0.000000
```

Now I will explain how to find out the dissimilarity or distance matrix; import pandas as pd from scipy.spatial import distance_matrix. The data equal to 1, 4; 2, 5; 3, 6 so there are three dataset for two variables. If you want to know the, the distance between a and b; now we have three dataset for two variables. If you want to know the distance matrix, so pd.DataFrame(distance_matrix(df.values, df.values)), so you will get this is distance matrix. So 3 variables and 3 dataset. So this, the distance matrix between 0 and 0 is 0; between 1 and 0 is 1.41, between 2 and 0 is 2.82.

(Refer Slide Time: 16:42)

```

jupyter Proximity_Measure Last Checkpoint: 9 minutes ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3.0
H 10 78

In [4]: M Distance_matrix = pd.DataFrame(distance_matrix(df.values), index=df.index, columns=df.index)
Distance_matrix

Out[4]:
          A   B   C   D   E   F   G   H
A  0.00000 69.83521 2.00000 71.589158 108.577462 95.710337 5.831662 17.720045
B  69.83521 0.00000 70.830784 5.656954 42.190348 26.248809 75.665730 87.209651
C  2.00000 70.690794 0.00000 72.446683 108.772324 96.706780 5.086600 17.262677
D  71.589158 42.190348 0.00000 41.963424 26.400708 77.388630 89.157168
E  108.577462 26.248809 70.877204 43.852424 0.00000 19.104973 114.337221 126.000000
F  95.710337 26.248809 96.706780 26.400708 19.104973 0.00000 101.548018 112.871858
G  5.831662 75.665730 5.086600 77.388630 114.337221 101.548018 0.00000 12.165525
H  17.720045 87.209651 17.262677 89.157168 126.000000 112.871858 12.165525 0.000000

```

Now, we will go for a distance calculation for that import pandas as pd import numpy as np, I will run it, so data matrix is given like this. So there are two variables. There are 8 persons. A, B, C, D, E, F, G, H. Suppose if you want to know the distance matrix so pd.DataFrame distance_matrix you follow this command, the distance matrix is this one, right between A and A is 0; between B and A is 69.83; between C and A it is 2.00.

(Refer Slide Time: 17:22)

```

jupyter Proximity_Measure Last Checkpoint: 9 minutes ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3.0
H 10 78

In [5]: M Distance_matrix.round(decimals=1, out=None)

Out[5]:
          A   B   C   D   E   F   G   H
A  0.0 69.8 2.0 71.6 108.6 96.7 5.6 17.7
B  69.8 0.0 70.8 57.4 42.2 26.2 75.7 87.2
C  2.0 70.8 0.0 72.4 108.9 96.8 5.1 17.3
D  71.6 6.7 72.4 0.0 41.9 26.4 77.4 89.2
E  108.6 42.2 109.9 45.9 0.0 19.1 114.3 128.0
F  95.7 26.2 98.0 26.4 19.1 0.0 101.5 112.9
G  5.8 75.7 51.7 74.4 114.3 101.5 0.0 12.2
H  17.7 87.2 17.0 69.2 125.0 112.9 12.2 0.0

```

So if you want to have 1 decimal with the rounding of 1 decimal we got this distance matrix. So this distance matrix is given as input for cluster analysis. In this lecture I have explained how to find out the dissimilarity matrix for mixed type of variables. What is the meaning of mixed type

of variables? When we got for a cluster analysis the dataset may be combination ordinal, interval and ratio data.

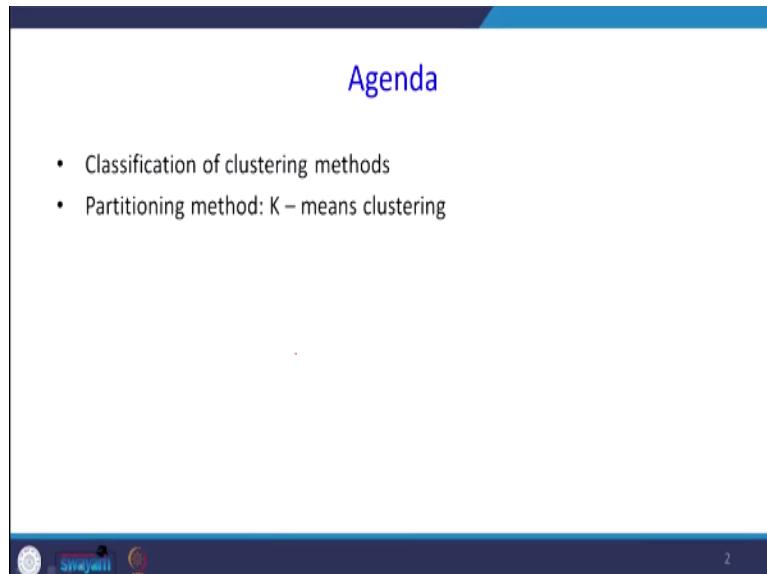
When these three types of data come together how to find out the dissimilarity matrix? That I have explained. Then, by using python I have explained how to find out the distance. So I have explained how to find out the Euclidean distance. Then I have explained how to find out the Manhattan distance and then I have explained how to find the Minkowski distance. At the end I have explained with the help of python how to find out the distance matrix for the interval-scaled data, because that distance matrix can be used as a input for our cluster analysis. Thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 54
K-Means Clustering

In this lecture, we will talk about K – means Clustering. Before that, I will explain what are the classifications of this clustering method. There are two type of classifications that I will explain in this class. In one classification is a K - means Clustering that I will solve one problem numerically with the help of some example. After solving the problem numerically, I will go to python there I will explain how to use python for doing this K - means clustering.

(Refer Slide Time: 00:59)



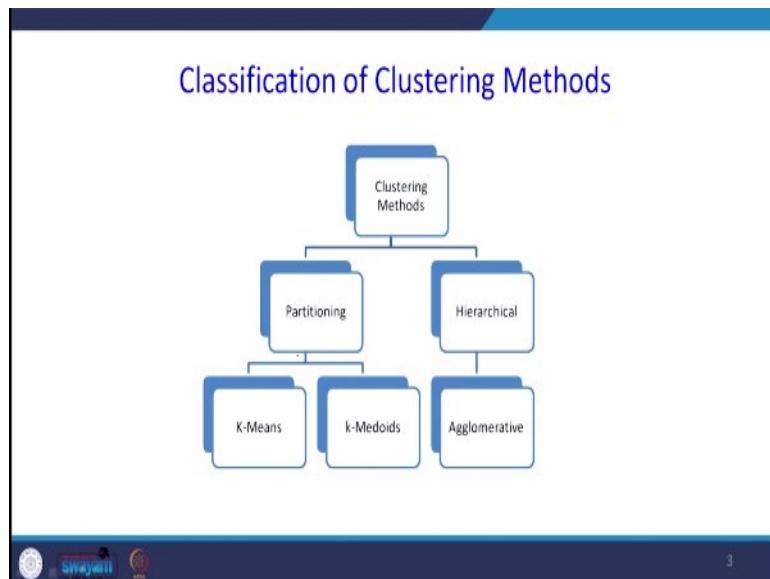
The slide has a dark blue header and footer bar. The title 'Agenda' is centered in the header. The main content area contains a bulleted list:

- Classification of clustering methods
- Partitioning method: K – means clustering

In the footer bar, there are three small circular icons on the left and the number '2' on the right.

So the agenda for this lecture is classification of clustering methods under which the partitioning method K – means clustering. That we will see in this class.

(Refer Slide Time: 01:03)



So this picture shows the classification of clustering methods. So the clustering methods generally classified into two category, one is partitioning method another one is hierarchical method. In partitioning there is another classifications one is K-Means another one is K-Medoids. In hierarchical there are two methods, one is Agglomerative method another method is Divisive method that will explain when I am explaining this hierarchical method. So in this lecture we are going to discuss about K-Means clustering.

(Refer Slide Time: 01:39)

Which Clustering Algorithm to Choose

- The choice of a clustering algorithm depends on
 - Type of data available
 - Particular purpose
- It is permissible to try several algorithms on the same data, because cluster analysis is mostly used as a descriptive or exploratory tool



So which clustering algorithm to choose? Because previously I was saying two method one is see the partitioning another one is hierarchical. The K-Means algorithm is generally used in advance if you know how many clustering is required. That time you can go for this partitioning

method. If you do not have idea how much cluster you need to do then you can go for hierarchical. So the another point, the choice of clustering algorithm depends upon type of data available and particular purpose.

Particular purpose in the sense whether you want to have in advance how many cluster is required or let us go for all type of classifications later we will give to user to chose the right number of clustering. It is permissible to try several algorithm on the same data, because cluster analysis is mostly used as a descriptive or exploratory tool.

(Refer Slide Time: 02:37)

Partitioning Method

Given -

- a data set of n objects
- k , the number of clusters
- A partitioning algorithm organizes the objects into k partitions ($k \leq n$), where each partition represents a cluster.
- The clusters are formed to optimize an objective partitioning criterion
- Objective partitioning criterion such as a dissimilarity function based on distance
- Therefore, the objects within a cluster are "similar," whereas the objects of different clusters are "dissimilar" in terms of the data set attributes.

5

First we will talk about partitioning method. In partitioning method what are the data which are given is, a data set of n objects and k ; this is user-defined, k is a number of clusters. In advanced we are going to know how many cluster we are going to have. A partitioning algorithm organizes the objects into k partitions where $k \leq n$, where each partition represents a cluster. The clusters are formed to optimize an objective partitioning criterion.

I will explain what the partitioning criterion is in next slide. The objective partitioning criterion such as dissimilarity function based on distance. So what is happened, within the cluster the dissimilarity should very less between the cluster the dissimilarity should be more. Therefore, the objects within the cluster are similar, whereas the objects of different clusters are dissimilar in terms of dataset attributes.

(Refer Slide Time: 03:40)

Partitioning Method

- Partitioning methods are applied if one wants to classify the objects into k clusters, where k is fixed.



6

So partitioning methods are applied if one wants to classify the objects into k clusters, where k is fixed.

(Refer Slide Time: 03:49)

K-Means Method

- It is a centroid based technique
- The k -means algorithm takes the input parameter, k , and partitions a set of n objects into k clusters
- So that the resulting intra-cluster similarity is high but the inter-cluster similarity is low
- Cluster similarity is measured in regard to the *mean* value of the objects in a cluster, which can be viewed as the cluster's *centroid* or *center of gravity*



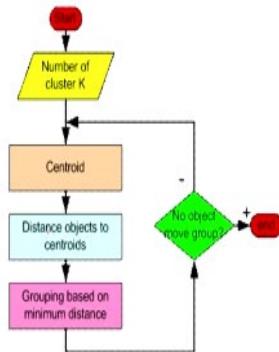
7

It is a centroid based technique because the; the centroid is nothing your mean, here kind of center of gravity. The k -means algorithm takes the input parameter, k , and partitions a set of n objects into k clusters, because as I told you here the k is one of input parameter. So, that the resulting intra-cluster similarity is high but inter-cluster similarity is; so what is happening is suppose there is a cluster 1 and cluster 2 so within that clusters there is a highly homogenous that the inter-cluster similarity is very low.

But between this cluster there should be a low similarity that means the, the dissimilarity is high. So cluster similarity is measured in regard of mean value of the objects in the cluster, which can be viewed as the clusters centroid or center of gravity as I told you.

(Refer Slide Time: 04:45)

Working Principle of K-Means Algorithm



So working principle of K-Means algorithm is; this is flowchart start, in advance you should number of clusters. Then you form the centroids. Randomly you can choose certain point then you form the centroids. Then the distance objects to the centroids. You find out suppose there is object, how far away that object is from the centroid. Then grouping based on the minimum distance. If there are two points, we have to take the point which is closed to that centroid into that cluster. Now that you have to continue for all points, if no object move the group then you can stop it otherwise you continue this cycle.

(Refer Slide Time: 05:32)

Working Principle of K-Means Algorithm

- First it randomly selects k of the objects, each of which initially represents a cluster mean or center
- For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean
- It then computes the new mean for each cluster
- This process iterates until the criterion function converges



9

Working principle of K-Means algorithm. First, it randomly selects k set of objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the objects and the cluster mean. It then computes the new mean for each cluster. Here what I mean, mean is this centroid. This process iterates until the criterion function converges.

(Refer Slide Time: 06:05)

Working Principle of K-Means Algorithm

- Criterion function

$$E = \sum_{i=1}^k \sum_{p \in C_i} |p - m_i|^2$$

where

- E is the sum of the square error for all objects in the data set;
- p is the point in space representing a given object;
- m_i is the mean of cluster C_i (both p and m_i are multidimensional).
- For each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed.
- This criterion tries to make the resulting k clusters as compact and as separate as possible.



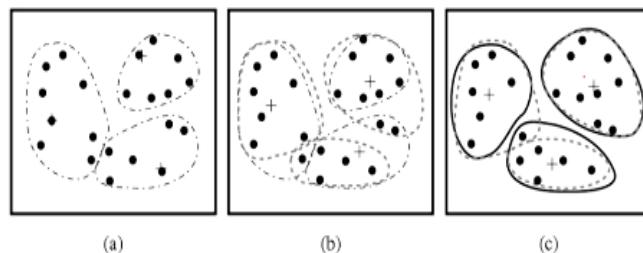
10

Let us see what is this criterion function. For example, $E = \sum_{i=1}^k \sum_{p \in C_i} |p - m_i|^2$ that means odd number clusters. P for all clusters. So the $|p - m_i|^2$ modulus squared values. Here what is the p is the point in space representing the given object, n is the mean of the cluster. It is nothing but kind of your mean absolute deviation but we are squaring that absolute deviation then we are

summing for all clusters. For each objects in each cluster, the distance from the object to its cluster center is squared, and the distance are summed. This criterion tries to make the resulting k clusters as compact as separate as possible.

(Refer Slide Time: 06:59)

K = 3



For example, this is $k = 3$, suppose we; there are n type of dataset. Randomly, I am making 3 clustering. After finding 3 clustering then I finding centroid of each clusters then from each centroid I am looking at all other objects and their distance then; if any point is closer to that centroid I am bringing that point into that cluster. Then I am updating this cluster and so on and continuing until all the objects are grouped into 3 clusters and the intra-distance is less and inter-distance is more. This I will explain with the help of a numerical example.

(Refer Slide Time: 07:45)

K-Means Clustering Algorithm

Algorithm: k-means. The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

- Input:
 - k: the number of clusters,
 - D: a data set containing n objects.
- Output: A set of k clusters.

12

Algorithm, k-means. The k-means algorithm for partitioning, where each cluster's center represented by the mean value of the object in the cluster. What are the input for the k-means algorithm? The number of clusters and data set containing n objects. What is going to be output? A set of k clusters.

(Refer Slide Time: 08:05)

K-Means Clustering Algorithm

Algorithm: k-means. The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

- Input:
 - k: the number of clusters,
 - D: a data set containing n objects.
- Output: A set of k clusters.

12

The K-Means clustering methods, as I; this also I have explained in my; that flowchart. Arbitrarily choose k objects from D from the dataset as the initial cluster centers. Repeat for all the points. Assign each object to the cluster to which the object is most similar. The distance is very small based on the mean value of the object in the cluster. Update the cluster means, i.e., calculate the mean value of the objects for each cluster until no change.

(Refer Slide Time: 08:40)

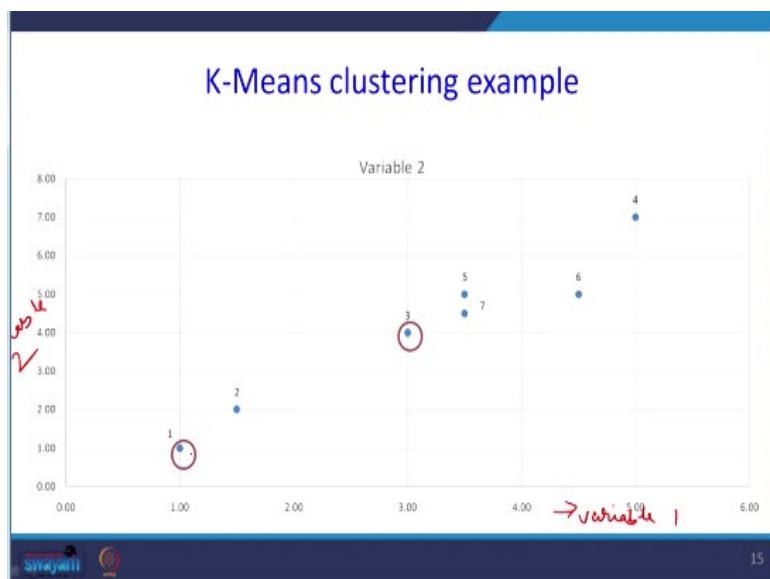
K-Means clustering example

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

14

Now taking a numerical example. There are 7 individuals. 1, 2, 3, 4, 5, 6, 7. There are two variables, variable 1 and variable 2. As I told you there is a small difference between clustering and factor analysis. In factor analysis we will group the variable into different category but in the cluster analysis the respondents the individuals have to group. That is the difference. Now there are 7 people is there. We are going to cluster this 7 people into some numbers. Let us see what is that number.

(Refer Slide Time: 09:17)



15

Suppose here the $k = 2$ initially here assume that the k is given $k = 2$, I want to make 2 clustering. Suppose randomly I have chosen, in x-axis the variable 1, in y-axis the variable 2. So

the point 1 and 3 are randomly taken, yes it is mentioned variable 2 and variable 1. Point 1 and 3 are taken randomly so there are $k = 2$, 2 cluster.

(Refer Slide Time: 10:05)

K-Means clustering example

- Initialization: Randomly we choose following two centroids ($k=2$) for two clusters. In this case the 2 centroid are:

Cluster	Var1	Var2
K1	1.0	1.0
K2	3.0	4.0

- Calculate Euclidean distance using the given equation

$$\text{Distance } [(x_1, y_1), (x_2, y_2)] = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

So the point 1 and 3 say the position of point 1 is 1, 2, the position of point 3 is 3, 4. So what is the initialization? Randomly we choose following two centroids where $k = 2$ for two clusters. In this case the two centroids are that point itself, K1 (1, 1); K2, (3, 4). So calculate the Euclidean distance using this given equations between all the points and between the cluster, so the formula for finding the Euclidean distance is root of (($x_2 - x_1$) whole square + ($y_2 - y_1$) whole square).

(Refer Slide Time: 10:47)

K-Means clustering example

Distance of k1 from k1 (1.0, 1.0) = $\sqrt{(1.0 - 1.0)^2 + (1.0 - 1.0)^2} = 0$

k1 to k2 (1.0, 1.0), (3.0, 4.0) = $\sqrt{(3.0 - 1.0)^2 + (4.0 - 1.0)^2} = 3.61$

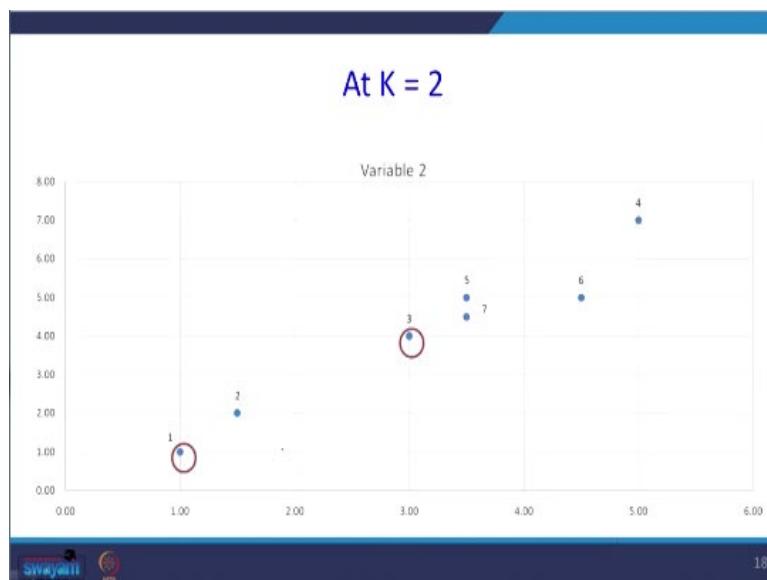
Distance of k2 from k2 (3.0, 4.0) = $\sqrt{(3.0 - 3.0)^2 + (4.0 - 4.0)^2} = 0$

Cluster	Centroid		
	K1	K2	Assignment
K1	0	3.61	k1
K2	3.61	0	k2

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

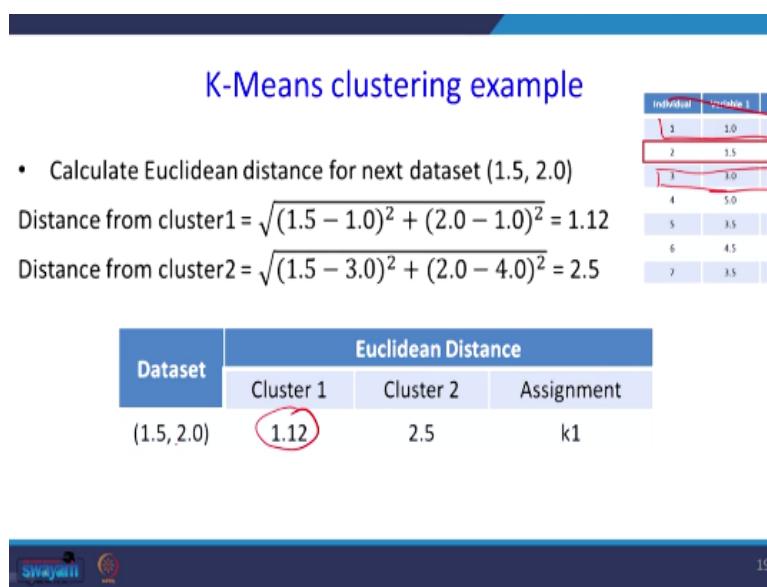
Since there are 2 points, there are two things we are going to do that. The distance between k1 and k1 from that point itself what was the distance from that same point, so the distance is 0. And between the two clusters that is 1 and 3 that is K1 and K2. The position of K1 is 1, 1, the position of K2 is 3, 4. The distance is $\sqrt{3^2 + 1^2} = \sqrt{10} \approx 3.16$. Then the distance of K2 from K2 itself, obviously that will be 0. So what I have taken cluster K1 and K2, the distance between two clusters, so the distance between K1 and K2 is 3.16 the same value is this one. Now we will update this.

(Refer Slide Time: 11:39)



So there are two parts, this is Cluster 1 and Cluster 2.

(Refer Slide Time: 11:45)



Now what we are going to do, we will take next variable 1.; that is the individual 2, 1.5 in 2. So from Cluster 1 I am going to find out how much faraway this point 2. Similarly, from cluster 2 that is individual 3, what is the distance of 2. So the distance from cluster 1 it is see $(1.5 - 1)$ whole square + $(2 - 1)$ whole square. So the distance from cluster 2 because 1and 3 is initial cluster, you have to remember this.

This was our initial cluster. So that is $(1.5 - 3)$ whole square + $(2 - 4)$ whole square. So the Euclidean distance the data set is 1.5, 2. The distance between cluster 1 and this data set is 1.12. The distance between cluster 2 and the data set is 2.5. So this point is closer to cluster 1 because it is the distance is less 1.1. So what we are going to do we are going to assign this point that is individual into the cluster 1 that is the K1. So this point is assigned to K1.

(Refer Slide Time: 13:05)



So what happened, this point is 1.5, 2. So now in this cluster 2 and 1 are assigned into same cluster. After assigning we are going to update the centroid of this cluster that is point 1 and 2.

(Refer Slide Time: 13:24)

K-Means clustering example

- Update the cluster centroid

Cluster	Var1	Var2
K1	$(1.0 + 1.5)/2 = 1.25$	$(1.0 + 2.0)/2 = 1.5$
K2	3.0	4.0

21

Now we will update the centroid of that cluster K1 because, why we are updating that K1 initially 1 individual now one more individual has entered into that cluster K1 so we are updating for that attributes, so the centroid of K1 is; for variable 1, it is $1 + 1.5$ divided by 2 it is 1.25. Then for variable 2 the centroid is in K1 for variable 2 the centroid is $1 + 2$ divided by 2 1.5. The K2 remain as it is.

(Refer Slide Time: 13:57)

K-Means clustering example

- Calculate Euclidean distance for next dataset (5.0, 7.0)

$$\text{Distance from cluster1} = \sqrt{(5.0 - 1.25)^2 + (7.0 - 1.5)^2} = 6.66$$

$$\text{Distance from cluster2} = \sqrt{(5.0 - 3.0)^2 + (7.0 - 4.0)^2} = 3.61$$

Individual	Variable 1	Var2
1	1.0	
2	1.5	
3	3.0	
4	5.0	
5	3.5	
6	4.5	
7	3.5	

Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	Assignment
(5.0, 7.0)	6.66	3.61	K-2

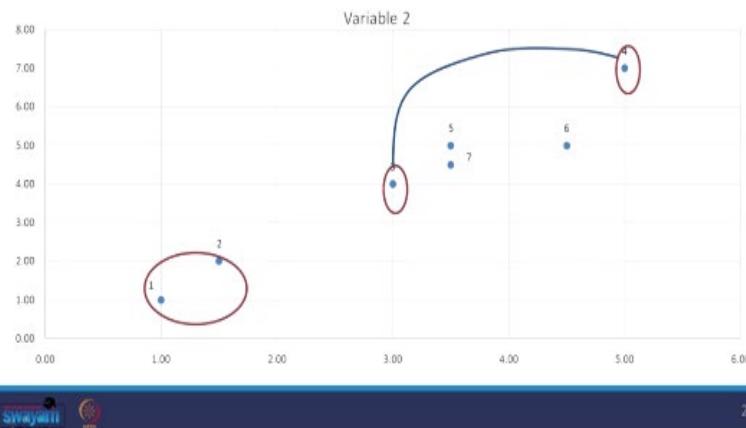
22

Now we will look at individual 4 whose attribute are 5, 7. From cluster 1 let us find out how much distance it is. Similarly, from cluster 3 we will find out how much distance it is. But in cluster 1 already there are two point has come. So that the our centroid has been already updated. So the distance from cluster 1 is 5, see that this, this 1.25 you got from this value, it is a centroid

so this value centroid value for variable 1; $5 - 1.25$ whole square similarly $7 -$ this was the new centroid of cluster 1, you see that this is 1.5. So that square is 6.66.

From cluster 2 the distance is $5 - 3$ whole square + $7 - 4$ whole square, the distance is 3.61. So this value we brought in the table format. So that data set is 5, 7 from cluster 1 the distance is 6.66, from cluster 2 the distance is 3.61. So this point is very close to cluster 2, so we are going to assign this point to the cluster 2 so 5, 7.

(Refer Slide Time: 15:16)



Now what happened this point is very close to this one. So we assigned this into this cluster.

(Refer Slide Time: 15:25)

K-Means clustering example

- Update the cluster centroid

Cluster	Var1	Var2
K1	1.25	1.5
K2	$(3.0 + 5.0)/2 = 4$	$(4.0 + 7.0)/2 = 5.5$

So after assigning as I told you, we have to update the centroid of cluster 2 now, because cluster 2 initially we had only one point, now one more point is entered. So the new centroid is $3 + 5$ divided by 2 that is 4 for variable 1, for variable 2 the centroid is $4 + 7$ divided by 2 = 5.5. Now this is the 4 and 5.5 is the new centroid for K2.

(Refer Slide Time: 15:54)

K-Means clustering example

- Calculate Euclidean distance for next dataset (3.5, 5.0)

Distance from cluster1 = $\sqrt{(3.5 - 1.25)^2 + (5.0 - 1.5)^2} = 4.16$

Distance from cluster2 = $\sqrt{(3.5 - 4.0)^2 + (5.0 - 5.5)^2} = 0.71$

Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	Assignment
(3.5, 5.0)	4.16	0.71	K-2

swayam IITM

25

Now we will add another variable 5, that is 3.5. We will find out how far away this point or this individual from cluster 1 and cluster 2. First we will find out from cluster 1. For a cluster 1 it is a $(3.5 - \text{the centroid of cluster 1})$ whole square + $(5 - \text{centroid of that is } 1.5)$ whole square that is a 4.16. Now this point $(3.5 - 4)$, 4 is centroid of our cluster 2, see this one $(3.5 - 4)$ whole square + $(5 - 5.5)$ this 5.5 you got from this updated centroid of cluster 2, so 0.71.

So let us bring this value into the table format. So the distance between this point and the cluster 1 is 4.16 and the distance between 3.5, 5 this point 2 cluster 2 is 0.71. So this point is this dataset is closer to the cluster 2, so we will assign this point also to cluster 2. So after assigning what has happened, so this point is assigned to cluster 2.

(Refer Slide Time: 17:07)

K-Means clustering example

- Update the cluster centroid

Cluster	Var1	Var2
K1	1.25	1.5
K2	$(3.0+5.0+3.5)/3 = 3.83$	$(4.0+7.0+5.0)/3 = 5.33$

27

Now we will go to the next variable. After assigning before going to the next variable we will update the centroid of cluster 2, because in cluster 2 now there are three points is there, that is 3, 4, 5, 7, 3.5, 5. First in variable 1 we will find out centroid nothing but the average 3.83 for, in K2 the centroid of variable 2 is 5.33. Now this is the new centroid for our K2.

(Refer Slide Time: 17:40)

K-Means clustering example

- Calculate Euclidean distance for next dataset (4.5, 5.0)

$$\text{Distance from cluster1} = \sqrt{(4.5 - 1.25)^2 + (5.0 - 1.5)^2} = 4.78$$

$$\text{Distance from cluster2} = \sqrt{(4.5 - 3.83)^2 + (5.0 - 5.33)^2} = 0.75$$

Individual	Variable 1	Var2
1	1.0	
2	1.5	
3	3.0	
4	5.0	
5	3.5	
6	4.5	
7	3.5	

Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	Assignment
(4.5, 5.0)	4.78	0.75	K- 2

28

Now we will take new individual and whose data point is 4.5 and 5. Now let us see how much distance or how much away from cluster 2. From cluster 1 say $4.5 - 1.25$ whole square + $(5 - 1.5)$ whole square 4.78, that is a 4.78. Now from cluster 2 let us see how much distance. $4.5 - 3.83$, how we got 3.83, because we are updated the centroid of cluster 2, so 3.83 whole square + $5 - 5.33$. How we got 5.33? This value, 5.33 whole squares so distance is

this one. By looking at the table this dataset is closer to the cluster 2 so we will assign this dataset also to K2.

(Refer Slide Time: 18:35)



So after assigning, what is happening, so this point is assigning to the cluster 2. Again we will update.

(Refer Slide Time: 18:41)

K-Means clustering example

- Update the cluster centroid

Cluster	Var1	Var2
K1	1.25	1.5
K2	$(3.0+5.0+3.5+4.5)/4= 4.00$	$(4.0+7.0+5.0+5.0)/4= 5.25$

Now in the cluster 2, there are 4 dataset, that is (3, 4) (5, 7) (3.5, 5) (4.5, 5). Now we will find the centroids. There are four dataset add it divided 4 that is 4, here 4.0+7.0+5.0 divided 4 that is 5.25, this is our centroid of K2, updated centroid.

(Refer Slide Time: 19:11)

K-Means clustering example

- Calculate Euclidean distance for next dataset (3.5, 4.5)

$$\text{Distance from cluster1} = \sqrt{(3.5 - 1.25)^2 + (4.5 - 1.5)^2} = 3.75$$

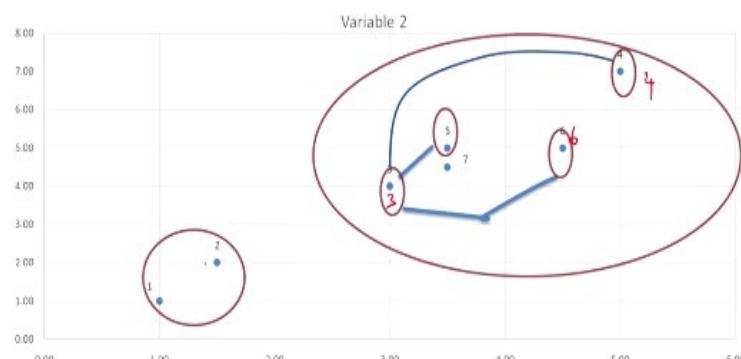
$$\text{Distance from cluster2} = \sqrt{(3.5 - 4.00)^2 + (4.5 - 5.25)^2} = 0.86$$

Individual	Variable 1	V
1	1.0	
2	1.5	
3	3.0	
4	5.0	
5	3.5	
6	4.5	
7	3.5	

Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	Assignment
(3.5, 4.5)	3.75	0.86	K-2

Now we will take the last point, it is the 3.5 and the 4.5. Let us see how much distance. This point is from cluster 1, cluster 2. From cluster 1, $3.5 - 1.25$ whole square + $4.5 - 1.5$ whole square that is 3.75. From cluster 2, how we got this 4 from this value $3.5 - 4$ whole square + $4.5 - 5.25$, so this value is -5.25 whole square is 0.86. So that value is 0.86. Again, so this point is closer to the cluster 2 so we will assign this point into cluster 2.

(Refer Slide Time: 19:50)



So now we have assigned, so this is one cluster. What are the point in this cluster? This is 3, 5, 7, 6, 4. In another cluster it is 1 and 2.

(Refer Slide Time: 20:06)

K-Means clustering example

- Update the cluster centroid

Cluster	Var1	Var2
K1	1.25	1.5
K2	$(3.0+5.0+3.5+4.5+3.5)/5 = 3.9$	$(4.0+7.0+5.0+5.0+4.5)/5 = 5.1$

33

After that again we will find out centroid of that cluster 2. So 1, 2, 3, 4, 5 dataset so add all the value divided by 5 it is 3.9. Again you add all the value divided by 5 it is 5.1. Now there are two clusters. The centroid of cluster 1 is 1.25, 1.5. The centroid of cluster 2 is 3.9 and 5.1. This value will verify when I am showing python demo.

(Refer Slide Time: 20:38)

K-Means clustering example

Individual	Variable 1	Variable 2	Assignment
1	1.0	1.0	1
2	1.5	2.0	1
3	3.0	4.0	2
4	5.0	7.0	2
5	3.5	5.0	2
6	4.5	5.0	2
7	3.5	4.5	2

34

Now this is the summary of our result. What has happened? So these individual is one group, cluster 1, this people in cluster 2. So what is the property is, the people in this cluster are more similar. People in this cluster also more similar. But between these two clusters the distance is far away. Now I have solved this problem with manually. Now we will go to python environment. So the same problem I will explain.

(Refer Slide Time: 21:23)

Python code for K- Means Clustering

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

In [2]: data = pd.read_excel('clustering_ex.xlsx')

In [3]: data
Out[3]:
   Variable_1  Variable_2
0         1.0        1.0
1         1.5        2.0
2         3.0        4.0
3         5.0        7.0
4         3.5        5.0
5         4.5        5.0
6         3.5        4.5
```

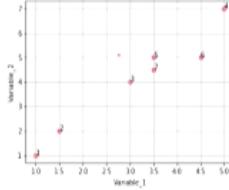
35

So I have brought this screenshot of our, for the K-Means clustering. We imported this required library import pandas as pd, import numpy as np, import matplotlib.pyplot as plt, so data is this one. So this was our data.

(Refer Slide Time: 21:41)

Python code for K- Means Clustering

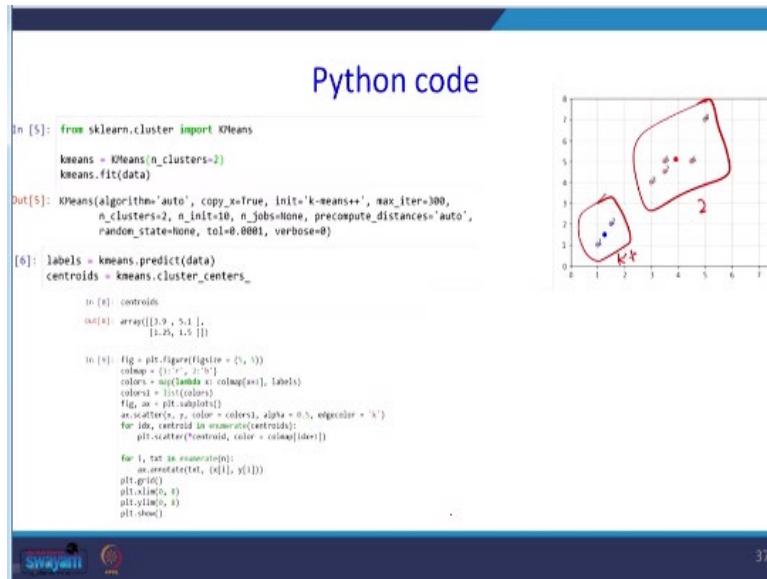
```
n [4]: fig = plt.figure(figsize = (5, 5))
x = data['Variable_1']
y = data['Variable_2']
n = range(1,8)
fig, ax = plt.subplots()
ax.scatter(x, y, marker='o', c='red', alpha=0.5)
plt.grid()
plt.xlabel("Variable_1")
plt.ylabel("Variable_2")
for i, txt in enumerate(n):
    ax.annotate(txt, (x[i], y[i]))
<matplotlib.figure.Figure at 0x20d7a5044a8>
```



36

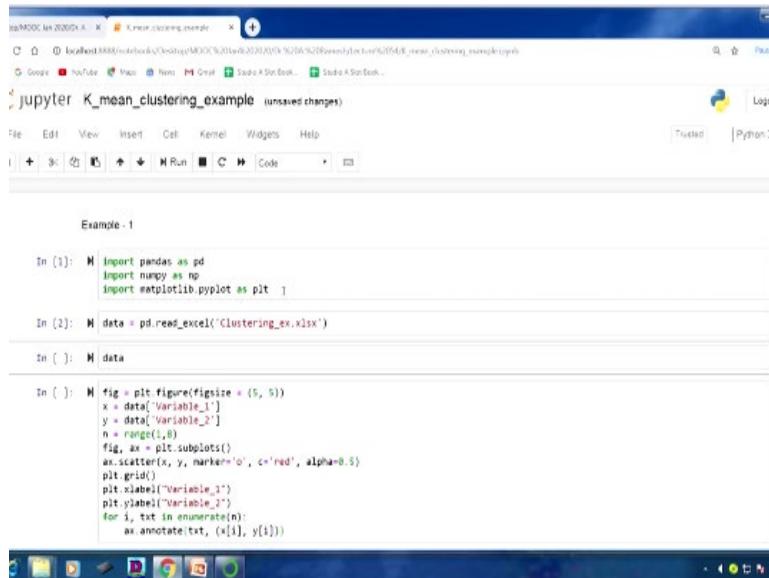
First we have plotted the scattered plot, the scattered plot with label.

(Refer Slide Time: 21:48)



Now this was the final result of cluster analysis. What happened? See this is one group, this is another group. This blue represents the centroid of this cluster 1, this red represents centroid of cluster 2. Now we will go to python environment. I will tell you how to do this K-means clustering in python.

(Refer Slide Time: 22:12)



Now I am going to explain how to use python for doing k-means algorithm. I have taken two examples; one example is what I have explained in my presentation. First we will import necessary libraries pandas, numpy, matplotlib.pyplot and so on. Next we will import the data.

(Refer Slide Time: 22:35)

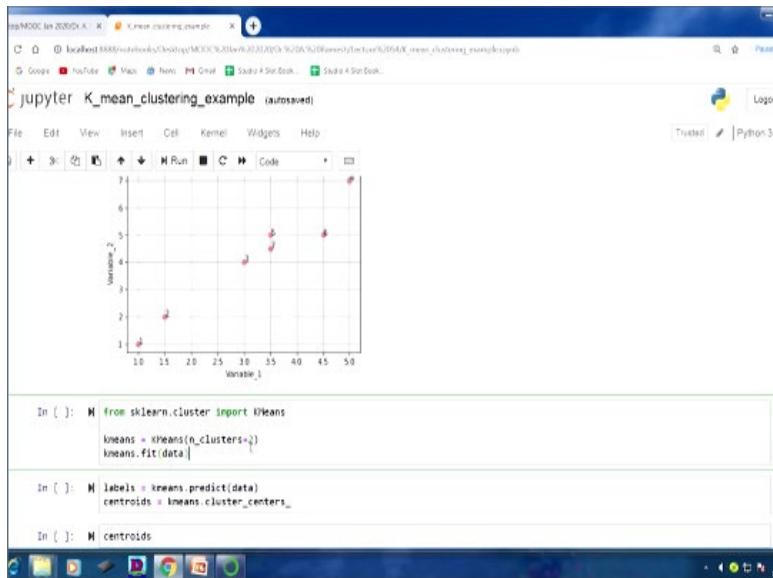
The screenshot shows a Jupyter Notebook interface. In the code cell, data is loaded from a CSV file and plotted as a scatter plot with two axes: Variable_1 (x-axis) and Variable_2 (y-axis). The data points are red circles.

```
In [1]: fig = plt.figure(figsize=(5, 5))
x = data['Variable_1']
y = data['Variable_2']
n = range(1,8)
fig, ax = plt.subplots()
ax.scatter(x, y, marker='o', c='red', alpha=0.5)
plt.grid()
plt.xlabel("Variable_1")
plt.ylabel("Variable_2")
for i, txt in enumerate(n):
    ax.annotate(txt, (x[i], y[i]))
```

```
In [2]: from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=2)
kmeans.fit(data)
```

The data, when you look at the data there are 7 individual is there, variable 1 and variable 2. After that we will position these individuals into, in a two-dimensional graph.

(Refer Slide Time: 22:50)



So what is happening, now we are able to see that all individuals, there are 7 individuals and their position, for example, the position of individual 1 is 1, 1, for position of individual 2 is 1.5, 2 and so on. For running k-means clustering algorithm we have to import this library. From sklearn.cluster import KMeans, so kmeans = KMeans(n_clusters = 2). So this is k = 2. If you want to have 3 clusters that is our next example, you have to substitute in sub 2 = 3, we will run this.

(Refer Slide Time: 23:27)

```

In [5]: KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
n_clusters=2, n_init=10, n_jobs=None, precompute_distances='auto',
random_state=None, tol=0.0001, verbose=0)

Out[5]: KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
n_clusters=2, n_init=10, n_jobs=None, precompute_distances='auto',
random_state=None, tol=0.0001, verbose=0)

In [6]: labels = kmeans.predict(data)
centroids = kmeans.cluster_centers_

In [7]: centroids
Out[7]: array([3.9, 5.1],
[1.25, 1.5])

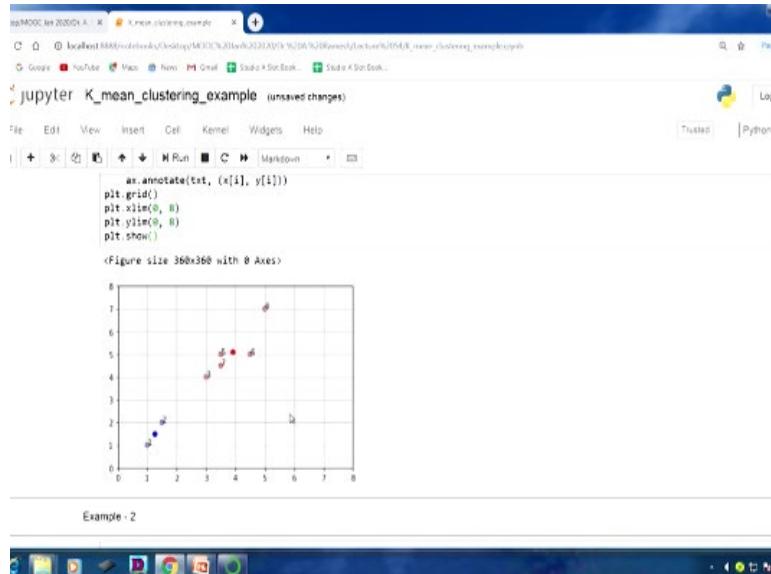
In [8]: fig = plt.figure(figsize=(5, 5))
colmap = {1: 'r', 2: 'b'}
colors = np.array([colmap[x+1] for x in labels])
colorsl = list(colors)
fig, ax = plt.subplots()
ax.scatter(x, y, color=colormap, alpha=0.5, edgecolor='k')
for idx, centroid in enumerate(centroids):
    plt.scatter(*centroid, color=colormap[idx+1])

for i, txt in enumerate(n):
    ax.annotate(txt, (x[i], y[i]))
plt.grid()
plt.xlim(0, 8)
plt.ylim(0, 8)

```

After running, we will verify, now the two clusters has been formed. Now we will verify the centroid of the two clusters. So the centroid of the two clusters is 3.9, 5.1 that was my cluster 2 centroid. For cluster 1 the centroid is 1.25, 1.5.

(Refer Slide Time: 23:50)



Let us see in picture form. This is the final output. This blue says, this is a centroid of cluster 1. Here the red one says that this is centroid of cluster 2. So what happening now two clusters are formed, in cluster 1 individual 1 and 2 is there; in cluster 2 individual 3, 5, 7, 6, 4 is there. This was exactly the result which I have done in the presentation.

(Refer Slide Time: 24:21)

```

In [9]: data = pd.read_excel('datapoints.xlsx')
        data
Out[9]:
   x  y
0  2  10
1  2  5
2  8  4
3  5  8
4  7  6
5  6  4
6  1  2
7  4  9

```

```

In [1]: fig = plt.figure(figsize=(5, 5))
X = data['x']
Y = data['y']

n = range(1, 8)
fig, ax = plt.subplots()
ax.scatter(X, Y, color = "red")

```

I will take another example where instead of $k = 2$ will go for 3 clusters, this is a different data set. So in that there are 8 individual is there. There are x variable and y variable.

(Refer Slide Time: 24:37)

```

In [9]: data
Out[9]:
   x  y
0  2  4
1  1  2
2  4  9

```

```

In [1]: fig = plt.figure(figsize=(5, 5))
X = data['x']
Y = data['y']

n = range(1, 8)
fig, ax = plt.subplots()
ax.scatter(X, Y, color = "red")
plt.grid()
plt.xlabel("x")
plt.ylabel("y")
for i, txt in enumerate(n):
    ax.annotate(txt, (X[i], Y[i]))

```

```

In [2]: kmeans = KMeans(n_clusters=3)
kmeans.fit(data)

```

```

In [3]: labels = kmeans.predict(data)
centroids = kmeans.cluster_centers_

```

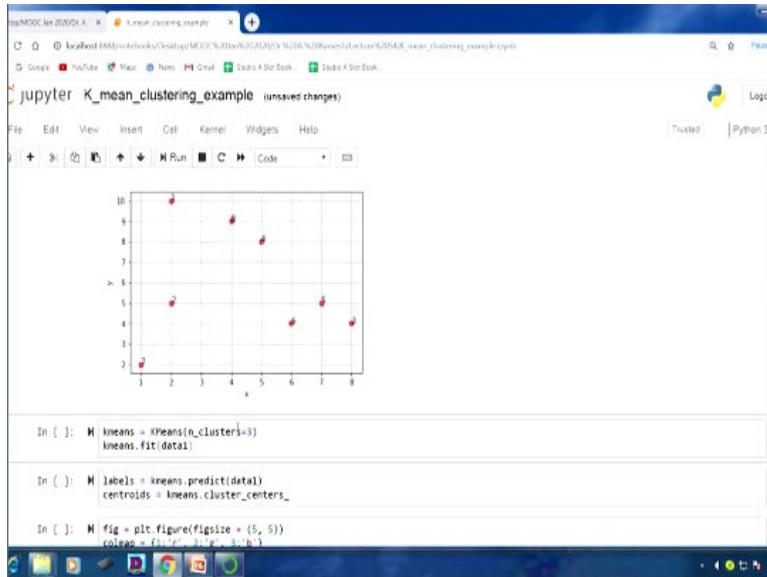
```

In [4]: fig = plt.figure(figsize=(5, 5))

```

The same way will plot into the two-dimensional plot, this was that way.

(Refer Slide Time: 24:41)



So there are 8 individual and their position.

(Refer Slide Time: 24:47)

```
In [11]: kmeans = KMeans(n_clusters=3, init='random', max_iter=300,
n_init=10, n_jobs=None, precompute_distances='auto',
random_state=None, tol=0.0001, verbose=0)

In [12]: labels = kmeans.predict(data)
centroids = kmeans.cluster_centers_

In [13]: centroids
```

```
Out[13]: array([[7.        , 4.33333333],
 [3.66666667, 9.        ],
 [1.5        , 3.5        ]])
```

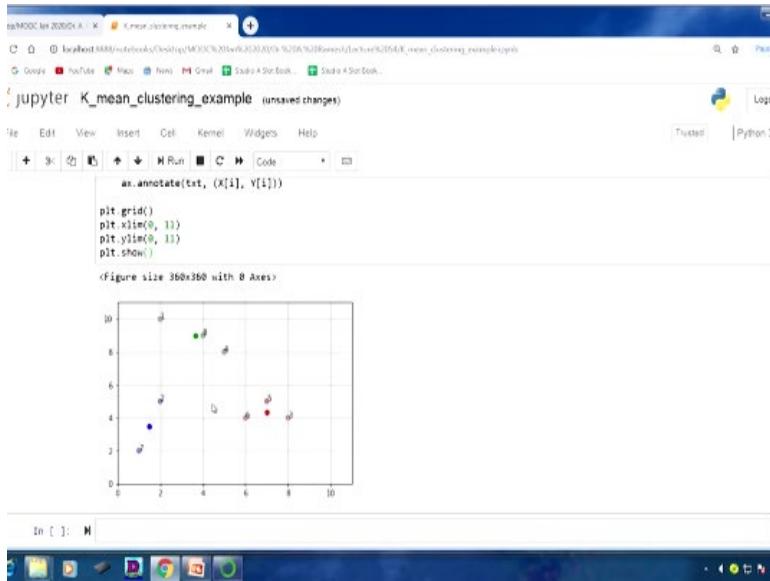
```
In [ ]: fig = plt.figure(figsize=(5, 5))
colmap = {1: 'r', 2: 'g', 3: 'b'}
colors = map(lambda X: colmap[X[1]], labels)
colorst = list(colors)
fig, ax = plt.subplots()
ax.scatter(X, Y, color=colors, alpha=0.5, edgecolor='k')
for idx, centroid in enumerate(centroids):
    plt.scatter(*centroid, color=colmap[idx])

for i, txt in enumerate(t):
    ax.annotate(txt, (X[i], Y[i]))

plt.grid()
```

We will go for k-means algorithm. You see that $k = 3$, probably we are going to have 3 clusters. So when you run that after running we can find out the centroid, so what I am going to do I am going to enter b then I go to show what is the value of centroid of these three clusters, so paste it, now you run it. So now there are three clusters. The centroid of this cluster is (7, 4.3) (3.6, 9) (1.5, 3.5).

(Refer Slide Time: 25:23)



Now I showed the picture from the final output, now just show the final output. There are three clusters which are in different color. So this blue says the centroid of cluster 1, this red says the centroid of cluster 2, this green says centroid of cluster 3. In this lecture, I have explained the classification of clustering methods. We know that there are two types of classification one is partitioning method another one is hierarchical method.

In the partitioning method there are another two classification one is a K-means clustering algorithm another one is K-Medoids. In hierarchical also there are two classification one is agglomerative and divisive method. But in this lecture I have covered only k-means algorithm. I have taken one numerical problems with the help of that numerical problems I have explained step-by-step procedure how to go for k-means algorithm.

After that I have explained the same problem in python, how to make k-means algorithm where $k = 2$. Apart from that, I have taken one more example in python environment then there I have explained how to make three clusters by taking the value of $k = 3$. The lecture I will explain the agglomerative method of clustering with the help of an example. Thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology - Roorkee

Lecture – 55
Hierarchical Method of Clustering - I

In our previous lecture, I have explained about K - means algorithm that is one type of clustering technique.

(Refer Slide Time: 00:32)

Agenda

- Introduction to Hierarchical clustering
- Partitioning Vs. Hierarchical

There is another type of technique is hierarchical method of clustering, let us see what is hierarchical method of clustering in this lecture and we will compare that partitioning method versus hierarchical clustering methods in this lecture. So, the agenda for this lecture is introduction to hierarchical clustering, then comparison of partitioning versus hierarchical clustering methods.

(Refer Slide Time: 00:53)

Introduction

- A hierarchical method creates a hierarchical decomposition of the given set of data objects
- A hierarchical clustering method works by grouping data objects into a tree of clusters
- A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed
- The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group

A hierarchical method creates a hierarchical decomposition of the given set of data objects, a hierarchical clustering method works by grouping data objects into tree of clusters. Here the tree of clusters, I will explain what is this tree of clusters in next slides. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed.

There are 2 way we can say in hierarchical method, one is agglomerative, the second one is divisive, the agglomerative approach also called bottom up approach, start with each object forming a separate group.

(Refer Slide Time: 01:42)

Introduction

- It successively merges the objects or groups that are close to one another, until all of the groups are merged into one (the topmost level of the hierarchy), or until a termination condition holds
- The divisive approach, also called the top-down approach, starts with all of the objects in the same cluster
- In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds

It successively merges the objects or groups that are close to one another, until all of the groups are merged into the top most level of the hierarchy or until a termination condition holds, on the other hand, the another classification in the hierarchical method is in divisive approach also called top down approach starts with of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster.

(Refer Slide Time: 02:27)

Introduction

- Hierarchical methods suffer from the fact that once a step (merge or split) is done, it can never be undone
- This rigidity is useful in that it leads to smaller computation costs by not having to worry about a combinatorial number of different choices
- However, such techniques cannot correct erroneous decisions

Or until a termination condition hold, hierarchical methods suffer from the fact that once a step is done, it can never be undone, so the problem with the hierarchical cluster is once a step is done, you cannot go back and correct the mistake. This rigidity is useful in that it leads to smaller computation cost by not having to worry about combinatorial number of different choices, however, such techniques cannot correct erroneous decisions that is only drawback of this hierarchical methods.

(Refer Slide Time: 03:04)

Agglomerative and Divisive Hierarchical Clustering

Agglomerative



- This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied
- Most hierarchical clustering methods belong to this category

Divisive Hierarchical



- This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster
- It subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions, such as a desired number of clusters is obtained or the diameter of each cluster is within a certain threshold

Let us compare agglomerative method versus divisive method, both are hierarchical method; let us see how this it is differ from each other. The agglomerative method this is bottom up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters until all of the objects are in a single cluster or until certain termination condition is satisfied.

Most hierarchical clustering methods belong to this category, on the other hand, the divisive method is a top down strategy does the reverse of agglomerative hierarchical clustering by starting with all object in one cluster. So, in divisive methods what we are doing, so you start from a bigger cluster, then you make smaller one, like it is a cutting a big cake into small pieces. On the other hand, the agglomerative is each object is separate clusters.

Then from that you can form all possible types of clusters that is kind of a tree, it is up to you to decide where you need to have the termination condition. In the divisive method, the divisive method sub divides the cluster into smaller and smaller pieces until each object forms a cluster on its own or until it satisfy certain termination condition such as a desired number of cluster is obtained or the diameter of each cluster is within the certain threshold.

(Refer Slide Time: 04:47)

Agglomerative versus divisive hierarchical clustering

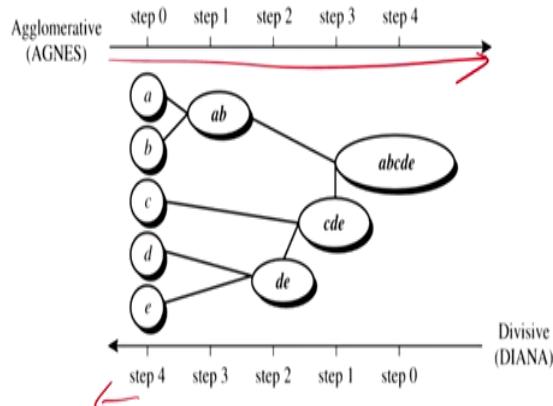


Figure: 1 Agglomerative and divisive hierarchical clustering on data objects{a,b,c,d,e}

This picture explains the difference between agglomerative and divisive method, you see that the arrow for agglomerative method is going on this side, it says that there are a, b, c, d, there are 5 objects, we start with each objects are in a separate cluster, then this a and b in step 0, all are clusters; 1, 2, 3, 4, 5 cluster is there, each cluster having only 1 unit. In step 1, see a and b are clubbed that is ab.

In step 2, d and e are clubbed, in step 3 this c, d, e are clubbed, in step 4 all these a, b, c, d is clubbed, so this is going in from left to right that is agglomerative method whereas in the divisive method, you start from; you see that look at this arrow it is going this side, start from the a, b, c, d that is a big by considering all the elements. A step 0; look at the step 0, only one cluster, in step 1, the c, d, e is 1, in step 2 de is another cluster. In step 3, from a, b, c, d again the ab has come out, in step 4 all individual elements are separate clusters that is a basic difference between agglomerative versus divisive hierarchical clustering.

(Refer Slide Time: 06:20)

Interpretation

- Figure: 1 shows the application of AGNES (AGglomerative NESting), an agglomerative hierarchical clustering method, and DIANA (DIvisive ANALysis), a divisive hierarchical clustering method, to a data set of five objects, {a,b,c,d,e}
- Initially, AGNES places each object into a cluster of its own
- The clusters are then merged step-by-step according to some criterion
- Let's say for example, clusters C_1 and C_2 may be merged if an object in C_1 and an object in C_2 form the minimum Euclidean distance between any two objects from different clusters

What is the interpretation of the previous slides? Figure 1 shows that application of agglomerative AGNES; agglomerative nesting, an agglomerative hierarchical clustering methods and DIANA divisive analysis, a divisive hierarchical clustering method to a data set of 5 objects; a, b, c, d, e. Initially, agglomerative method places each object into a cluster of its own, there is only one item. The clusters are then merged step by step according to some criterion, let us say for example, cluster C1 and C2 may be merged if an object C1 and the object C2 from the minimum Euclidean between any 2 objects from different clusters.

(Refer Slide Time: 07:13)

Interpretation

- This is a single-linkage approach in that each cluster is represented by all of the objects in the cluster, and the similarity between two clusters is measured by the similarity of the closest pair of data points belonging to different clusters
- The cluster merging process repeats until all of the objects are eventually merged to form one cluster

This is a single linkage approach, in that each cluster is represented by all of the objects in the cluster and the similarity between 2 cluster is measured by the similarity of the closest pair of the

data points belonging to different clusters. The cluster merging process repeats until all of the objects are eventually merged into form 1 cluster. So, what is happening here it is start from the step 0, it goes up to step 4, you see in step 0, there are 1, 2, 3, 4, 5 clusters in step 0.

(Refer Slide Time: 07:56)

Interpretation

- In DIANA, all of the objects are used to form one initial cluster
- The cluster is split according to some principle, such as the maximum Euclidean distance between the closest neighboring objects in the cluster
- The cluster splitting process repeats until, eventually, each new cluster contains only a single object
- In either agglomerative or divisive hierarchical clustering, the user can specify the desired number of clusters as a termination condition

But in step 1, all are merged into only 1 cluster that is a, b, c, d, e. In DIANA that is in the divisive method, all of the objects are used to form 1 initial cluster, this cluster split according to some principle such as maximum Euclidean distance between the closest neighbouring objects in the cluster. The cluster splitting process repeats until eventually each new cluster contains only 1 object, only a single object.

In either agglomerative or divisive hierarchical clustering, the user can specify the desired number of clusters as a termination condition. So, here the explanation of divisive method is start from here, in step 0 only 1 cluster is there. In step 4, there are 5 clusters is there, that is a difference.

(Refer Slide Time: 08:43)

Dendrogram

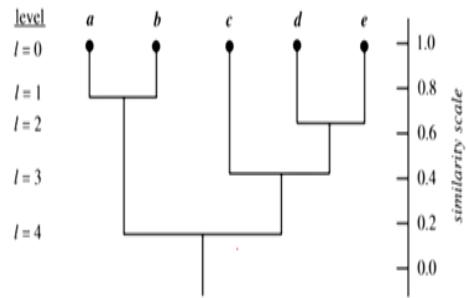


Figure 2: Dendrogram representation for hierarchical clustering of data objects{a,b,c,d,e}

In hierarchical clustering, another important terminology which you have to understand is dendrogram. What is a dendrogram? It is a kind of a tree kind of structure and it says there are different levels, level 0, 1, 2, 3, 4 on left hand side and right hand side there is a similarity scale. At level 0, there are 1, 2, 3, 4, 5 clusters a, b, c, d, e all are forming its own cluster. In level 1, ab is forming 1 cluster.

In level 2, the position of c is compared, in the position of c we are finding the distance between c and between these cluster a and b and cluster d and e. If it is closer to d and e, then c, d, e form an another cluster that is a level 3. Level 4, it is only 1 cluster, all are 5 elements are present in there, so this kind of picture is called dendrogram.

(Refer Slide Time: 09:45)

Dendrogram

- A tree structure called a dendrogram is commonly used to represent the process of hierarchical clustering
- It shows how objects are grouped together step by step
- Figure: 2 shows a dendrogram for the five objects presented in Figure:1 , where $l=0$ shows the five objects as singleton clusters at level 0
- At $l=1$, objects a and b are grouped together to form the first cluster, and they stay together at all subsequent levels

Dendrogram, a tree structure called a dendrogram is commonly used to represent the process of hierarchical clustering; it shows how objects are grouped together step by step. Figure 2 shows a dendrogram for the 5 objects presented in the figure 1, where l equal to 0, at level 0 shows the 5 objects are singleton clusters, there is only 1 element in the clusters at level 0. At level 1, object a and b are grouped together to form the first cluster and they stay together at all subsequent levels, this hierarchical structure can be understood with the help of this dendrogram.

(Refer Slide Time: 10:27)

Dendrogram

- We can also use a vertical axis to show the similarity scale between clusters
- For example, when the similarity of two groups of objects, $\{a,b\}$ and $\{c,d,e\}$, is roughly 0.16, they are merged together to form a single cluster

We can also use a vertical axis to show the similarity scale between the clusters actually, it is given on the right hand side of the picture. For example, when the similarity of 2 groups of object a and b and c, d, e is roughly 0.16, so they merged together to form a single cluster.

(Refer Slide Time: 10:49)

Measures for distance between clusters

- Four widely used measures for distance between clusters are as follows, where $|p-p'|$ is the distance between two objects or points, p and p' , m_i is the mean for cluster, C_i and n_i is the number of objects in C_i
- Minimum distance: $d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p-p'|$
- Maximum distance: $d_{\max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} |p-p'|$
- Mean distance: $d_{\text{mean}}(C_i, C_j) = |m_i - m_j|$
- Average distance: $d_{\text{avg}}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p-p'|$

Now, let us go to another important idea of measures of distance between clusters, there are 4 widely used measures for distance between clusters are as follows, where modulus of $p - p'$ dash is the distance between 2 objects or points that is a p and p' dash, where m_i is the mean of the cluster, C_i and n_i is the number of objects in C_i . The first measure is minimum distance, the minimum distance between cluster C_i and C_j equal to, so minimum of modulus of $p - p'$ dash.

The maximum distance is d_{\max} between cluster i and j equal to maximum of modulus of $p - p'$ dash, the mean distance; there is another measure is d_{mean} is the modulus of mean of the 2 clusters. The average distance is 1 divided by $(n_i - n_j)$, here n represents number of object in cluster 1, then sigma of $p - p'$ dash modulus.

(Refer Slide Time: 12:08)

Measures for distance between clusters

- When an algorithm uses the minimum distance, $d_{\min}(C_i, C_j)$, to measure the distance between clusters, it is sometimes called a nearest-neighbor clustering algorithm
- Moreover, if the clustering process is terminated when the distance between nearest clusters exceeds an arbitrary threshold, it is called a single-linkage algorithm
- If we view the data points as nodes of a graph, with edges forming a path between the nodes in a cluster, then the merging of two clusters, C_i and C_j , corresponds to adding an edge between the nearest pair of nodes in C_i and C_j

When an algorithm uses the minimum distance that is a d_{\min} between C_i and C_j , that is a distance between C_i and C_j . To measure the distance between clusters, it is sometime called nearest neighbour clustering algorithm, I will show you in picture in coming slides. Moreover, if the clustering process is terminated, when the distance between the nearest clusters exceeds the arbitrary threshold, it is called single linkage algorithm.

If we view the data points as nodes of graph with edges forming a path between the nodes in a cluster, then the merging of 2 clusters; C_i and C_j corresponds to adding an edge between the nearest pair of nodes C_i and C_j .

(Refer Slide Time: 12:55)

Measures for distance between clusters

- Because edges linking clusters always go between distinct clusters, the resulting graph will generate a tree
- Thus, an agglomerative hierarchical clustering algorithm that uses the minimum distance measure is also called a minimal spanning tree algorithm
- When an algorithm uses the maximum distance, $d_{\max}(C_i, C_j)$, to measure the distance between clusters, it is sometimes called a farthest-neighbor clustering algorithm
- If the clustering process is terminated when the maximum distance between nearest clusters exceeds an arbitrary threshold, it is called a complete-linkage algorithm

Because edges linking clusters always go between distinct clusters, the resulting graph will generate a tree, thus an agglomerative hierarchical clustering algorithm that uses the minimum distance measure is also called minimal spanning tree algorithm, even in your subject operation research also, there is a topic network problems in that you might have studied minimal spanning tree algorithm.

When an algorithm uses the maximum distance between cluster i and j, to measure the distance between clusters, it is sometime called farthest neighbour clustering algorithm. If the clustering process is terminated, when the maximum distance between nearest cluster exceeds an arbitrary threshold, it is called complete linkage algorithm.

(Refer Slide Time: 13:57)

Measures for distance between clusters

- By viewing data points as nodes of a graph, with edges linking nodes, we can think of each cluster as a complete sub graph, that is, with edges connecting all of the nodes in the clusters
- The distance between two clusters is determined by the most distant nodes in the two clusters
- Farthest-neighbor algorithms tend to minimize the increase in diameter of the clusters at each iteration as little as possible
- If the true clusters are rather compact and approximately equal in size, the method will produce high-quality clusters
- Otherwise, the clusters produced can be meaningless

By viewing data points as nodes of graph with edges linking nodes, we can think of each cluster as a complete sub graph that is with edges connecting all of the nodes in the cluster. The distance between 2 cluster is determined by the most distant nodes in the 2 clusters, farthest neighbour algorithm tend to minimise the increase in diameter of the clusters at each iteration as little as possible. If the true clusters are rather compact and approximately equal in size, the method will produce high quality clusters, otherwise the clusters produced can be meaningless.

(Refer Slide Time: 14:40)

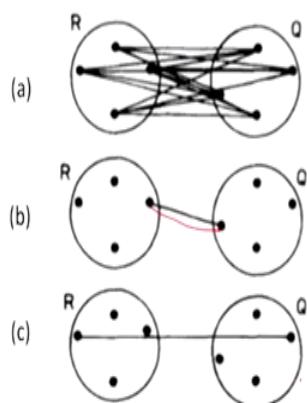
Choice of measurement

- The above minimum and maximum measures represent two extremes in measuring the distance between clusters
- They tend to be overly sensitive to outliers or noisy data
- The use of mean or average distance is a compromise between the minimum and maximum distances and overcomes the outlier sensitivity problem
- Whereas the mean distance is the simplest to compute, the average distance is advantageous in that it can handle categorical as well as numeric data

Let us go for choice of measurement; the above minimum and maximum measure represents 2 extremes in measuring the distance between clusters, they tend to be overly sensitive to outliers or noisy data. The use of mean or average distance is a compromise between minimum and maximum distances and overcome the outlier sensitivity problems, whereas the mean distance is the simplest to compute, the average distance is advantageous in that it can handle categorical as well as numeric data.

(Refer Slide Time: 15:24)

Illustration

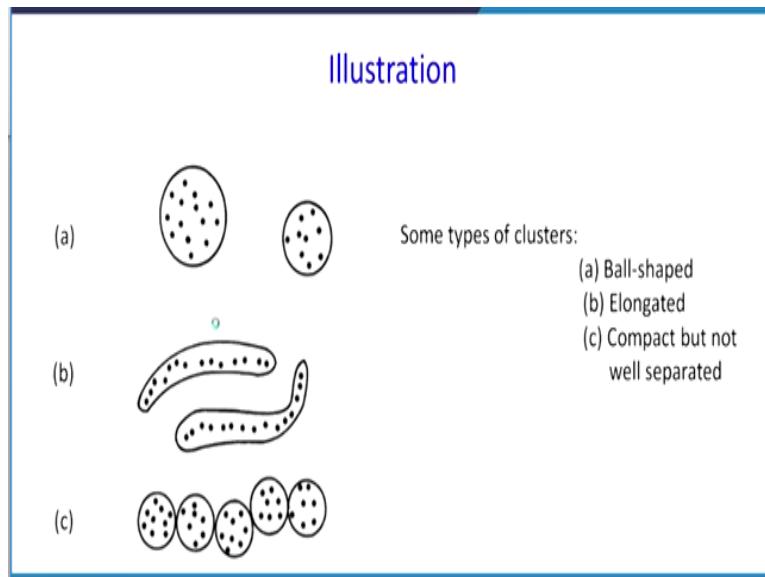


Representation of some definitions of inter-cluster dissimilarity: (a) Group average
(b) Nearest neighbor
(c) Furthest neighbor

This picture shows the distance measures, see the first one represents group average, so you see that all the points are connected with all other points in that cluster. This R is one cluster; Q is one cluster you see that we have finding the beverage that is a group average. This is

representation of some definition of inter cluster dissimilarity. The second one is the nearest neighbour. See, this is a nearest neighbour, the third one is the farthest neighbour that I have explained in my previous slides, this is a different type of distance measures.

(Refer Slide Time: 16:08)



So, this picture shows some type of clusters, see the cluster here is the ball shaped one, the second one is elongated one the last one is compact but not well separated. So, what will happen; if we follow the group average, your; the final cluster may be this shape, what is that; the ball shaped one. If we follow this distance measures that is a nearest neighbour, the final cluster may look like this one that is elongated.

You see that any time, it can form with this point any time we can go to that other cluster, in case if you follow the farthest neighbour distance measures, your final cluster may be in this format that is a compact but not well separated, that is why choosing the correct distance is more important, based on the your distance measures, your shape of final cluster also will vary.

(Refer Slide Time: 17:15)

Difficulties with hierarchical clustering

- The hierarchical clustering method, though simple, often encounters difficulties regarding the selection of merge or split points
- Such a decision is critical because once a group of objects is merged or split, the process at the next step will operate on the newly generated clusters
- It will neither undo what was done previously nor perform object swapping between clusters

The difficulties with the hierarchical clustering; the hierarchical clustering method, though simple often encounters difficulties regarding the selection of merge or split points, such a decision is critical because once a group of object is merged or split, the process at the next step will operate on the newly generated clusters. As I told you, this also one of the drawback, once the cluster is formed, you cannot, any mistake has happened that cannot be rectified if you follow hierarchical clustering methods. See it will neither undo what was done previously nor perform objects swapping between clusters; these are the some of the disadvantages of hierarchical clustering.

(Refer Slide Time: 18:02)

Difficulties with hierarchical clustering

- Thus merge or split decisions, if not well chosen at some step, may lead to low-quality clusters
- Moreover, the method does not scale well, because each decision to merge or split requires the examination and evaluation of a good number of objects or cluster
- For improving the clustering quality of hierarchical methods is to integrate hierarchical clustering with other clustering techniques, resulting in multiple-phase clustering

Thus, merge or split decisions if not well chosen at some step may lead to low quality clusters, moreover, the method does not scale well because each decision to merge or split requires the examination and the evaluation of good number of objects or clusters. For improving the cluster quality of hierarchical method is to integrate hierarchical clustering with other clustering techniques resulting in multiple phase clustering. So, what we can do; if you want to improve the quality of hierarchical clustering, it can be clubbed with other clustering algorithms, so that you can improve the quality of our clustering.

(Refer Slide Time: 18:49)

Partitioning Vs. Hierarchical

Method	General Characteristics
Partitioning methods	<ul style="list-style-type: none">- Find mutually exclusive clusters of spherical shape- Distance-based- May use mean or medoid (etc.) to represent cluster center- Effective for small- to medium-size data sets
Hierarchical methods	<ul style="list-style-type: none">- Clustering is a hierarchical decomposition (i.e., multiple levels)- Cannot correct erroneous merges or splits- May incorporate other techniques like microclustering or consider object "linkages"

Now, let us compare the partitioning clustering algorithm versus hierarchical clustering algorithm, first we will see what is this partitioning methods, what are the general characteristics that is a K means algorithm is a partitioning method. Find mutually exclusive cluster of spherical shape, this partitioning method is a distance based, may use mean or medoid to represent cluster center, effective for small to medium sized dataset.

Hierarchical methods; clustering is a hierarchical decomposition at multiple levels, cannot correct erroneous merges or splits, may incorporate other techniques like micro clustering, consumer object linkages.

(Refer Slide Time: 19:39)

K-means versus Hierarchical clustering

Let us compare K means versus hierarchical clustering.

(Refer Slide Time: 19:43)

K means versus hierarchical clustering

K- means clustering

- Non-hierarchical methods-(k-means), using a pre-specified number of clusters
- This method assigns records to each cluster to find the mutually exclusive cluster of spherical shape based on distance
- In this case, one can use mean or median as a cluster centre to represent each cluster

Hierarchical clustering

- Hierarchical methods can be either agglomerative or divisive
- Agglomerative methods begin with 'n' clusters and sequentially merge similar clusters until a single cluster is obtained

K means clustering; it is a non-hierarchical method because there will be K means using a pre specified number of clusters. So, when we doing K means clustering in advance we know, how many cluster we are going to have. This method assigns records to each cluster to find the mutually exclusive cluster of spherical shape based on distance. In K mean clustering one can use mean or median as a cluster centre to represent each cluster.

For hierarchical clustering, this method can be either agglomerative or divisive. Agglomerative method begins with n clusters and sequentially merge similar clusters until a single cluster is obtained.

(Refer Slide Time: 20:32)

K means versus hierarchical clustering	
K- means clustering	Hierarchical clustering
<ul style="list-style-type: none">This method is generally less computationally intensive and are therefore preferred with very large datasets	<ul style="list-style-type: none">Divisive methods work in the opposite direction, starting with one cluster that includes all recordsHierarchical methods are especially useful when the goal is to arrange the clusters into a natural hierarchy

K means clustering methods is generally less computationally intensive and are therefore preferred with very large data set. In hierarchical clustering, divisive methods work in the opposite direction starting with one cluster that includes all the records. Hierarchical methods are especially useful when the goal is to arrange the cluster into a natural hierarchy.

(Refer Slide Time: 21:01)

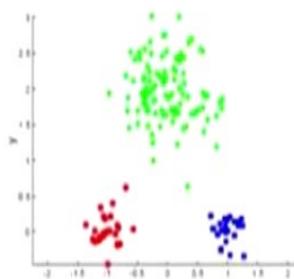
K means versus hierarchical clustering	
K- means clustering	Hierarchical clustering
<ul style="list-style-type: none">A partitioning (K- means) clustering a simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset)	<ul style="list-style-type: none">A hierarchical clustering is a set of nested clusters that are organized as a tree

A partitioning that means a K means clustering simply a division of the set of data objects into non overlapping subsets clusters such that each data object is in exactly one subset; a hierarchical clustering is a set of nested clusters that are organised as a tree.

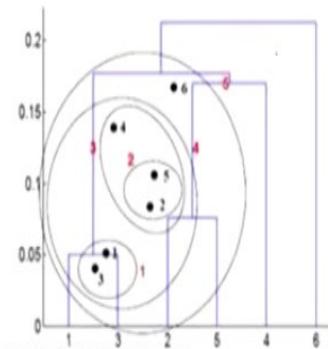
(Refer Slide Time: 21:23)

K means versus hierarchical clustering

Un-nested cluster



Nested cluster



Ashok, A.R., Prabhakar, C.R. and Dyaneshwar, P.A., Comparative Study on Hierarchical and Partitioning Data Mining Methods.

When you look at this picture, the picture shows in the left hand side is un-nested clusters, we can say it is a K means clusters. In the right hand side, the name is called nested cluster, this is nothing but your agglomerative or hierarchical clustering.

(Refer Slide Time: 21:41)

K means versus hierarchical clustering

- Hierarchical clustering does not assume a particular value of 'k', as needed by k-means clustering
- The generated tree may correspond to a meaningful taxonomy
- Only a distance or "proximity" matrix is needed to compute the hierarchical clustering

	a	b	c	d	e	f	
a	0	184	222	177	216	231	Proximity matrix
b	184	0	45	123	128	200	
c	222	45	0	129	121	203	
d	177	123	129	0	46	83	
e	216	128	121	46	0	83	
f	231	200	203	83	83	0	

Hierarchical clustering does not assume a particular value of K as needed by K means clustering, the generated tree may correspond to a meaningful taxonomy, only a distance or proximity

matrix is needed to compute the hierarchical clustering. This is an example of proximity matrix, see the between a and a, the distance is 0, proximity 0, between b and a, the distance is 184.

(Refer Slide Time: 22:12)

K means versus hierarchical clustering

K Means clustering

- In K Means clustering, since one start with random choice of clusters, the results produced by running the algorithm multiple times might differ
- K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D)

Hierarchical clustering

- Results are reproducible in Hierarchical clustering .
- Hierarchical clustering don't work as good as, k means when the shape of the clusters is hyper spherical

In K means clustering, since one start with random choice of clusters, the result produced by running the algorithm multiple times might differ. K means is found to work well, when the shape of the cluster hyper spherical like circle in 2 dimension, sphere in 3 dimension. In hierarchical clustering, results are reproducible, hierarchical clustering do not work as good as K means, when the shape of the cluster is hyper spherical.

(Refer Slide Time: 22:51)

K means versus hierarchical clustering

K Means clustering

- K Means clustering requires prior knowledge of K i.e. no. of clusters one want to divide your data into

Hierarchical clustering

- In hierarchical clustering one can stop at any number of clusters, one find appropriate by interpreting the dendrogram

So, the K means clustering suitable for hyper spherical clustering, K means clustering requires prior knowledge of K that is number of clusters one want to divide your data into. In hierarchical clustering, one can stop at any number of clusters, one find appropriate by integrating the dendrogram.

(Refer Slide Time: 23:04)

K means versus hierarchical clustering



There are 2 pictures, the top one is example of K means clustering where K equal to 3, the bottom one is hierarchical clustering, you see that there is a hierarchy is there, so this is an example of.

(Refer Slide Time: 23:18)

Hierarchical clustering

Advantages

- Ease of handling of any forms of similarity or distance
- Consequently, applicable to any attributes types

Advantage of hierarchical clustering; ease of handling of any form of similarity or distance, consequently applicable to any attribute types, here attribute is the variable types, it may be interval, it may be ratio, it may be binary or categorical.

(Refer Slide Time: 23:38)

Limitations of Hierarchical Clustering

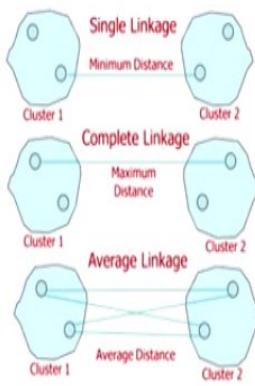
- Hierarchical clustering requires the computation and storage of an $n \times n$ distance matrix
- For very large datasets, this can be expensive and slow
- The hierarchical algorithm makes only one pass through the data
- This means that records that are allocated incorrectly early in the process cannot be reallocated subsequently
- Hierarchical clustering also tends to have low stability
- Reordering data or dropping a few records can lead to a different solution

Some of the limitations of hierarchical clustering; hierarchical clustering requires the computation and storage of n cross n distance matrix, here n is number of objects. For very large dataset this can be expensive and slow, the hierarchical algorithm makes only one pass through the data, this means that the records that are allocated incorrectly early in the process cannot be reallocated subsequently. Hierarchical clustering also tends to have low stability, reordering data or dropping a few records can lead to different solution.

(Refer Slide Time: 24:18)

Limitations of Hierarchical Clustering

- With respect to the choice of distance between clusters, single and complete linkage are robust to changes in the distance metric (e.g., Euclidean, statistical distance) as long as the relative ordering is kept.
- In contrast, average linkage is more influenced by the choice of distance metric, and might lead to completely different clusters when the metric is changed
- Hierarchical clustering is sensitive to outlier



Limitation of hierarchical clustering; with respect to the choice of distance between clusters, single and complete linkages are robust to changes in the distance metric as long as the relative order is kept. So, what is the example of single linkage when you look at this, there is a cluster 1, cluster 2, so the minimum distance is called single linkage and the distance between the farthest points that is called complete linkage.

So, if you use these distance measure, then the cluster what you got is very robust, in contrast average linkage is more influenced by choice of distance metrics and might lead to completely different clusters when the metric is changed, hierarchical clustering is sensitive to outlier. If any extreme dataset is there that may provide different kind of clusters.

(Refer Slide Time: 25:14)

Average-linkage clustering

- Compromise between Single and Complete Link
- Strengths
 - Less susceptible to noise and outliers
- Limitations
 - Biased towards globular clusters



Then, when we will go for average linkage clustering, what is an example of average linkage clustering? This one, you see that, all the distance are connected then we found average. It is a compromise between single and complete link. The strength of average linkage clustering is less susceptible to noise and outliers, the limitations are biased towards globular clusters. So, when you use average linkage clustering, many times, the cluster may be like a spherical shape.

(Refer Slide Time: 25:58)

K-means clustering

Advantages

- The center of mass can be found efficiently by finding the mean value of each co-ordinate
- This leads to an efficient algorithm to compute the new centroids with a single scan of the data

Disadvantages

- K-means has problems when clusters are of differing sizes, densities, non-globular shapes and when the data contains outliers

Now, let us see the advantage of K means clustering, previously we have seen advantage of hierarchical clustering, the advantage of K means clustering is the centre of mass can be found efficiently by finding the mean value of each coordinate, this leads to an efficient algorithm to compute the new centroids with a single scan of data. The disadvantages are K means has

problem when the cluster of different sizes, densities, non- globular shapes and when the data contains outliers.

(Refer Slide Time: 26:34)

Similarity

- Two most popular methods: hierarchical agglomerative clustering and k-means clustering
- In both cases, we need to define two types of distances: distance between two records and distance between two cluster
- In both cases, there is a variety of metrics that can be used

What is a similarity between hierarchical clustering and K means clustering; 2 most popular method is hierarchical agglomerative clustering and K means clustering, in both cases we need to define 2 types of distance, distance between 2 records and distance between 2 cluster, in both cases, there is a variety of metrics that can be used. In this lecture, I have explained introduction to hierarchical clustering.

Then, I have compared the difference between K means clustering techniques and hierarchical clustering techniques and also I have explained the advantages and disadvantages. In the next lecture, we are going to take one numerical example, with the help of numerical example; I am going to explain how to do a hierarchical clustering, thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology - Roorkee

Lecture – 56
Hierarchical Method of Clustering - II

In a previous lecture, I have explained introduction to hierarchical clustering and different types of distance measures. In this lecture, I have taken a numerical example, with the help of the numerical example, I am going to explain how to do hierarchical clustering method, for that same problem, I am going to explain how to use Python for doing hierarchical clustering.

(Refer Slide Time: 00:53)

Agenda

- Agglomerative hierarchical algorithm
- Python demo

So, the agenda for this lecture is agglomerative hierarchical algorithm, the second one is python demo.

(Refer Slide Time: 00:57)

Example for Hierarchical Agglomerative Clustering (HAC)

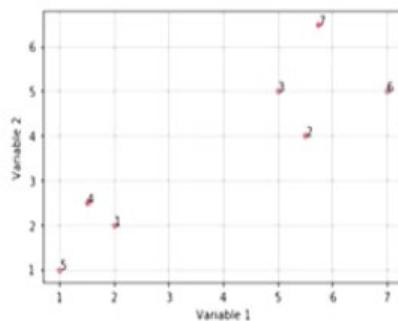
- A data set consisting of seven objects for which two variables were measured.

Object	Variable 1	Variable 2
1	2.00	2.00
2	5.50	4.00
3	5.00	5.00
4	1.50	2.50
5	1.00	1.00
6	7.00	5.00
7	5.75	6.50

So the example of hierarchical agglomerative clustering, HAC; a data set consists of 7 objects for which 2 variables were measured. There are 7 objects; 1, 2, 3, 4, 5, 6, 7, variable 1 and variable 2. So, variable 1 is 2, 5, 5.5, 5 and so on. Variable 2 is 2, 4, 5, 2, 1, 5, 6 and so on.

(Refer Slide Time: 01:24)

Scatter plot



When you plot in the scatterplot, it is appearing that there are 7 data set, variable 1 in x axis, variable 2 in y axis. Now, we are going to do hierarchical clustering for this data set.

(Refer Slide Time: 01:37)

Example for HAC

- Calculate Euclidean Distance and create the distance matrix.

$$\text{Distance}[(x_1, y_1), (x_2, y_2)] = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Distance (1,2)

$$(2.00, 2.00) (5.50, 4.00) = \sqrt{(5.50 - 2.00)^2 + (4.00 - 2.00)^2} = 4.02$$

Distance (1,3)

$$(2.00, 2.00) (5.00, 5.00) = \sqrt{(5.00 - 2.00)^2 + (5.00 - 2.00)^2} = 4.24$$

Distance (1,4)

$$(2.00, 2.00) (1.50, 2.50) = \sqrt{(1.50 - 2.00)^2 + (2.50 - 2.00)^2} = 0.71$$

Object	Var 1	Var 2
1	2.00	2.00
2	5.50	4.00
3	5.00	5.00
4	1.50	2.50
5	1.00	1.00
6	7.00	5.00
7	5.75	6.50

What is the first one; calculate the Euclidean distance and create a distance matrix, what we are going to do; first we are going to create a distance matrix, for that we are going to find out the distance between 1 and 2, 1 and 3, 1 and 4, 1 and 5, 1 and 6, 1 and 7 and 2 and 3, 2 and 4, 2 and 5, 2 and 6 and 2 and 7, then 3 and 4, 3 and 5, 3 and 6 and 3 and 7 then 4 and 5, 4 and 6 and 4 and 7, then 5 and 6 and 5 and 7 finally, 6 and 7.

First we will find the distance between object 1 and 2, so the object 1 is 2, 2, object 2 is 5.5, 4, after finding the Euclidean distance, we know the formula is $x_2 - x_1$ square + $y_2 - y_1$ whole square, then square root so, $5.5 - 2$ whole square + $4 - 2$ whole square, then square root, that is your 4.02. Then, let us find the distance between; now we have seen 1 and 2, now the distance between 1 and 3. So, 1 and 3 is the position of 1 is 2, 2, position of 3 is 5, 5, so it is $5 - 2$ whole square + $5 - 2$ whole square. So, it is 4.24, now let us find the distance between 1 and 4, so this 1 and 4. So, 2, 2 and 1.5, 2.5, so the distance is $1.5 - 2$ whole square + $2.5 - 2$ whole square that is 0.71.

(Refer Slide Time: 03:35)

Object	Var 1	Var 2
1	2.00	2.00
2	5.50	4.00
3	5.00	5.00
4	1.50	2.50
5	1.00	1.00
6	7.00	5.00
7	5.75	6.50

Example for HAC

Distance (1,5)

$$(2.00, 2.00) (1.00, 1.00) = \sqrt{(1.00 - 2.00)^2 + (1.00 - 2.00)^2} = 1.41$$

Distance (1,6)

$$(2.00, 2.00) (7.00, 5.00) = \sqrt{(7.00 - 2.00)^2 + (5.00 - 2.00)^2} = 5.83$$

Distance (1,7)

$$(2.00, 2.00) (5.75, 6.50) = \sqrt{(5.75 - 2.00)^2 + (6.50 - 2.00)^2} = 5.86$$

Now, we will find the distance between 1 and 5, so this 1 and 5 because this we have done it, so 1 and 5 is 2, 2; the position of the 5th object is 1, 1, so it is 1 – 2 whole square + 1 – 2 whole square, it is 1.41. Now, we will find the distance 1 and 6 that is 2, 2, then 7, 5, so it is 7 – 2 whole square + 5 – 2 whole square that is 5.83. Similarly, we can find the distance between 1 and 7 that is 2, 2 versus 5.75 and 6.5.

(Refer Slide Time: 04:30)

Object	Var 1	Var 2
1	2.00	2.00
2	5.50	4.00
3	5.00	5.00
4	1.50	2.50
5	1.00	1.00
6	7.00	5.00
7	5.75	6.50

Example for HAC

Distance (2,3)

$$(5.50, 4.00) (5.00, 5.00) = \sqrt{(5.00 - 5.50)^2 + (5.00 - 4.00)^2} = 1.12$$

Distance (2,4)

$$(5.50, 4.00) (1.50, 2.50) = \sqrt{(1.50 - 5.50)^2 + (2.50 - 4.00)^2} = 4.27$$

Distance (2,5)

$$(5.50, 4.00) (1.00, 1.00) = \sqrt{(1.00 - 5.50)^2 + (1.00 - 4.00)^2} = 5.41$$

Distance (2,6)

$$(5.50, 4.00) (7.00, 5.00) = \sqrt{(7.00 - 5.50)^2 + (5.00 - 4.00)^2} = 1.80$$

So, it is a 5.75 – 2 whole square + 6.5 – 2 whole square, we got 5.86, now we have done 1 versus all the point, now we go second versus 3, the distance between 2 and 3, the position of 2 is 5, 5, 4; 3 is 5, 5, so the distance is 5 – 5.5 whole square + 5 – 4 whole square, it is 1.12. Now, 2 and 4;

5.5, 4 is the position of 2, 4 is the position of 2, the position of object 4 is 1.5, 2.5, so the distance is $1.5 - 5.5$ whole square + $2.5 - 4$ whole square that is 4.27.

Next we will find the distance between 2 and 5; we know the position of 2 is 5.5, 4, the position of object 5 is 1, 1, see this point, the distance is $(1 - 5.5)$ whole square + $(1 - 4)$ whole square that is 5.41. Now, we can find the distance between 2 and 6, so the position of object 6 is 7, 5, so the distance is $(7 - 5.5)$ whole square + $(5 - 4)$ whole square that is 1.81.

(Refer Slide Time: 06:04)

Object	Var 1	Var 2
1	2.00	2.00
2	5.50	4.00
3	5.00	5.00
4	1.50	2.50
5	1.00	1.00
6	7.00	5.00
7	5.75	6.50

Example for HAC

Distance (2,7)

$$(5.50, 4.00) (5.75, 6.50) = \sqrt{(5.75 - 5.50)^2 + (6.50 - 4.00)^2} = 2.51$$

Distance (3,4)

$$(5.00, 5.00) (1.50, 2.50) = \sqrt{(1.50 - 5.00)^2 + (2.50 - 5.00)^2} = 4.30$$

Distance (3,5)

$$(5.00, 5.00) (1.00, 1.00) = \sqrt{(1.00 - 5.00)^2 + (1.00 - 5.00)^2} = 5.66$$

Distance (3,6)

$$(5.00, 5.00) (7.00, 5.00) = \sqrt{(7.00 - 5.00)^2 + (5.00 - 5.00)^2} = 2.00$$

The next one; we are going to find the distance between 2 and 7, so 2 and 7, so the position of object 7 is 5.75, 6.5, so the distance is $5.75 - 5.5$ whole square + $6.5 - 4$ whole square, so this difference whole square and this difference whole square that is 2.51. Now, 2; the point 2, the object 2 and other all 7's we compare, now we will compare 3; 3 and 4, 3 and 5, 3 and 6 and 3 and 7.

So, the 3 and 4 distance is the position of point '3' is 5, 5 and 4 is 1.5, 2.5, so this distance versus this distance, this difference, so $1.5 - 5$ whole square + $2.5 - 5$ whole square that is 4.30. Next 3 and 5, so 5, 5 and the 1; 1 is the position of object 5, so what will happen; $1 - 5$ whole square + $1 - 5$ whole square, 5.66. Now, 3 and 6; so 3 is 5, 5; 6 is 7, 5, so the difference is $7 - 5$ whole square + $5 - 5$ whole square, the distance is 2.

(Refer Slide Time: 07:49)

Example for HAC

Distance (3,7)

$$(5.00, 5.00) (5.75, 6.50) = \sqrt{(5.75 - 5.00)^2 + (6.50 - 5.00)^2} = 1.68$$

Distance (4,5)

$$(1.50, 2.50) (1.00, 1.00) = \sqrt{(1.00 - 1.50)^2 + (1.00 - 2.50)^2} = 1.58$$

Distance (4,6)

$$(1.50, 2.50) (7.00, 5.00) = \sqrt{(7.00 - 1.50)^2 + (5.00 - 2.50)^2} = 6.04$$

Distance (4,7)

$$(1.50, 2.50) (5.75, 6.50) = \sqrt{(5.75 - 1.50)^2 + (6.50 - 2.50)^2} = 5.84$$

Object	Var 1	Var 2
1	2.00	2.00
2	5.50	4.00
3	5.00	5.00
4	1.50	2.50
5	1.00	1.00
6	7.00	5.00
7	5.75	6.50

Now, we are going to find the next one; 3, 7, so 3, 7 what is the distance; we know position of 3 is 5, 5; 7 is 5.75 – 6.50, so this difference whole square versus this difference whole square, so 5.75 – 5 whole square + 6.5 – 5 whole square that is 1.68. Now, we have to find out 3 versus all the point, now we will take 4, 5 and 4, 6 and 4, 7. So, the distance between 4 and 5 is 1.5 – 2.5 is a the position of object 4, for the 5th one it is 1, 1.

When you look at the distance, it is 1.58, then 4 and 6 we have done that one, now we will go for 4 and 6. The 4 and 6 is the 4th 1.5 , 2.5, the 6th position is 7, 5, so the difference is 7 – 1.5 whole square + 5 – 2.5 whole square equal to 6.04. Now, we will find out 4 versus 7, the distance between 4 and 7, the 1.5, 2.5, the position of object 7 is 5.75 , 6.50, so the this difference whole square plus this difference whole square.

(Refer Slide Time: 09:25)

Example for HAC

Distance (5,6)

$$(1.00, 1.00) (7.00, 5.00) = \sqrt{(7.00 - 1.00)^2 + (5.00 - 1.00)^2} \\ = 7.21$$

Object	Var 1	Var 2
1	2.00	2.00
2	5.50	4.00
3	5.00	5.00
4	1.50	2.50
5	1.00	1.00
6	7.00	5.00
7	5.75	6.50

Distance (5,7)

$$(1.00, 1.00) (5.75, 6.50) = \sqrt{(5.75 - 1.00)^2 + (6.50 - 1.00)^2} = 7.27$$

Distance (6,7)

$$(7.00, 5.00) (5.75, 6.50) = \sqrt{(5.75 - 7.00)^2 + (6.50 - 5.00)^2} = 1.95$$

So, $5.5; 5.75 - 1.5$ whole square + $6.5 - 2.5$ whole square that is 5.84 , we have completed, now we will go 5 and 6 and 5 and 7, the distance between 5 and 6 is 1, 1 is a position of object 5, for 6th one, it is 7, 5, so it is $7 - 1$ whole square + $5 - 1$ whole square that is 7.21. Now, we will find out distance between 5 and 7 that is 1, 1 versus 5.75 – 6.5, so it is $5.75 - 1$ whole square + $6.5 - 1$ whole square, 7.27.

Now, 6 versus 7, we will find out the distance, so that is position of object 6 is 7, 5, position of object 7 is 5.75, 6.5, so the distance is $5.75 - 7$ whole square + $6.5 - 5.00$ whole square that is 1.95.

(Refer Slide Time: 10:24)

Distance Matrix

- The distance matrix is-

	1	2	3	4	5	6	7
1	0.0						
2	4.0	0.0					
3	4.2	1.1	0.0				
4	0.7	4.3	4.3	0.0			
5	1.4	5.4	5.7	1.6	0.0		
6	5.8	1.8	2.0	6.0	7.2	0.0	
7	5.9	2.5	1.7	5.8	7.3	2.0	0.0

Now, we have compared all the distance, as I told you this is a distance matrix, see that in the diagonal it is '0' because the distance is 0, we got distance between 2 and 1, 3 and 1, 4 and 1, 5 and 1, 6 and 1, 7 and 1, this value which we got from our previous slides, so we got the distance matrix.

(Refer Slide Time: 10:47)

Example for HAC

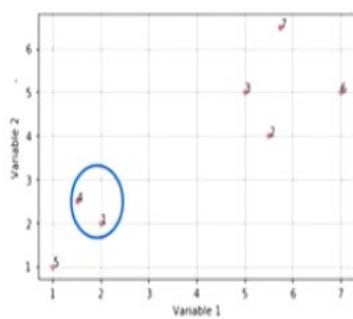
- Select minimum element to build first cluster formation-

	1	2	3	4	5	6	7
1	0.0						
2	4.0	0.0					
3	4.2	1.1	0.0				
4	0.7	4.3	4.3	0.0			
5	1.4	5.4	5.7	1.6	0.0		
6	5.8	1.8	2.0	6.0	7.2	0.0	
7	5.9	2.5	1.7	5.8	7.3	2.0	0.0

After the distance matrix, for that hierarchical agglomerative method, select minimum element to build the first cluster formation, so among this, find out where there is a minimum distance, so the minimum distance is this one; 0.7. What are the 2 objects between 4 and 1, so what is going to do; we are going to form a cluster, in that there are 2 element is going to be there; 4 and 1.

(Refer Slide Time: 11:23)

Example for HAC



So, the 4 and 1 will form a cluster, so we got this one, 4 and 1, this is our first cluster.

(Refer Slide Time: 11:29)

Example for HAC

- Recalculate distance to update distance matrix

$$\begin{aligned}
 & - \text{MIN}[\text{dist}(1,4), 2] = \text{MIN}(\text{dist}(1,2), (4,2)) \\
 & \quad = \text{MIN}(4.0, 4.3) = 4.0 \\
 & - \text{MIN}[\text{dist}(1,4), 3] = \text{MIN}(\text{dist}(1,3), (4,3)) \\
 & \quad = \text{MIN}(4.2, 4.3) = 4.2 \\
 & - \text{MIN}[\text{dist}(1,4), 5] = \text{MIN}(\text{dist}(1,5), (4,5)) = \text{MIN}(1.4, 1.6) = 1.4 \\
 & - \text{MIN}[\text{dist}(1,4), 6] = \text{MIN}(\text{dist}(1,6), (4,6)) = \text{MIN}(5.8, 6.0) = 5.8 \\
 & - \text{MIN}[\text{dist}(1,4), 7] = \text{MIN}(\text{dist}(1,7), (4,7)) = \text{MIN}(5.9, 5.8) = 5.8
 \end{aligned}$$

	1	2	3	4	5	6	7
1	0.0						
2	4.0	0.0					
3	4.2	1.1	0.0				
4	0.7	4.3	4.3	0.0			
5	1.4	5.4	5.7	1.6	0.0		
6	5.8	1.8	2.0	6.0	7.2	0.0	
7	5.9	2.5	1.7	5.8	7.3	2.0	0.0

We are going to find the distance between that cluster 1, 4 forming a distance versus 2, so what we are going to do; we are going to find out distance between 1 and 2 and 4 and 2, whichever is minimum, we are going to take that distance because in cluster 1, already there are 2 object is there, we are going to consider the minimum distance. So, the distance between 1, 2, see that this value 4 and 4, 2 is 4.3, this value we take from the table.

So, the minimum distance is 4, the next one 1, 4 versus 3, similarly 1, 4 versus 5, 1, 4 versus 6, 1, 4 versus 7, so if you want to know this cluster versus 3, so minimum distance between 1, 3 and 4, 3; this 1, 3 and 4, 3, so 1, 3 it is a 4.2; 4, 3 it is a 4.3, this value, so minimum is 4.2. Now, 4th we cannot go, already the 4th one is already gone to that cluster, so we will go to 5th, so 1, 4 versus 5th object, for that we have to find out the minimum distance between 1, 5 and 4, 5.

So, 1, 5 is 1.4, this value and 4, 5 is your this value, 1.6 that 1.6, the minimum is 1.4, then 1, 4 with 6, now what happened here; we have to find out the minimum distance 1, 6 and 4, 6. So 1, 6 is 5.8 this value and 4, 6 that is 6, this value, so minimum is 5.8. Now, why we are taking 1.4 because this we formed one cluster, so from this cluster there are 2 point; 2 object 1 and 4. From 1 and 4, this 7 is how far away?

So, 2 way we have to do; 1 and 7 we have to find the distance and 4 and 7 we have to find the distance whichever minimum that has to be kept. So, 1 and 7, 4 and 7, so 1 and 7 is this one, 5.9, 4 and 7 that is your 5.8, so minimum value is 4.8.

(Refer Slide Time: 14:26)

Example for HAC

- Updated distance matrix for the cluster (1, 4)

	1,4	2	3	5	6	7
1,4	0.0					
2	4.0	0.0				
3	4.2	1.1	0.0			
5	1.4	5.4	5.7	0.0		
6	5.8	1.8	2.0	7.2	0.0	
7	5.8	2.5	1.7	7.3	2.0	0.0

Now, we are going to update this value, so what update we have done that one; since 1 and 4 form a new cluster, so now we will find this value, how we got this value? So, 2, 1, 4; so 2, 1, 4 is 4, so we updated. Now, so updated the distance, this distance matrix is going to be used for further steps.

(Refer Slide Time: 14:50)

Example for HAC

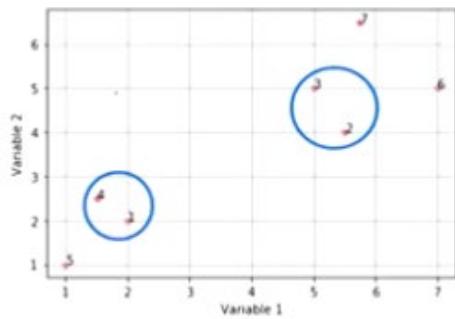
- Select minimum element to build next cluster formation-

	1,4	2	3	5	6	7
1,4	0.0					
2	4.0	0.0				
3	4.2	1.1	0.0			
5	1.4	5.4	5.7	0.0		
6	5.8	1.8	2.0	7.2	0.0	
7	5.8	2.5	1.7	7.3	2.0	0.0

So, now in that updated matrix, again you find out which is minimum, so this 1.1 is minimum that forms 3 and 2. So, now what is happening; the 3 and 2 is going to form one cluster.

(Refer Slide Time: 15:06)

Example for HAC



Yeah, 3 and 2 is formed one cluster.

(Refer Slide Time: 15:09)

Example for HAC

- Recalculate distance to update distance matrix
 - $\text{MIN}[\text{dist}(2,3), (1,4)] = \text{MIN}(\text{dist}(2,(1,4)), (3,(1,4))$
 $= \text{MIN}(4.0, 4.2) = 4.0$
 - $\text{MIN}[\text{dist}(2,3), 5] = \text{MIN}(\text{dist}(2,5), (3,5)) = \text{MIN}(5.4, 5.7) = 5.4$
 - $\text{MIN}[\text{dist}(2,3), 6] = \text{MIN}(\text{dist}(2,6), (3,6)) = \text{MIN}(1.8, 2.0) = 1.8$
 - $\text{MIN}[\text{dist}(2,3), 7] = \text{MIN}(\text{dist}(2,7), (3,7)) = \text{MIN}(2.5, 1.7) = 1.7$

	1,4	2	3	5	6	7
1,4	0.0					
2	4.0	0.0				
3	4.2	1.1	0.0			
5	1.4	5.4	5.7	0.0		
6	5.8	1.8	2.0	7.2	0.0	
7	5.8	2.5	1.7	7.3	2.0	0.0

Now, since 2 and 3 is formed a distance and already there is 1 cluster; 1, 4, so we are going to find out the distance between 2 and 1, 4 and 3 and 1, 4. So, 2 and 1, 4 you can find out this is 4 because 1 and 4 formed a cluster, then 3 and 1, 4; 3 and 1, 4 is 4.2, in this minimum is 4. Similarly, the distance between this newly formed cluster 2, 3 versus 5, 5th point, we are not go

to 4th one because 4 is already formed a cluster, so the 5th one, we have to find out the minimum distance of 2, 5 versus 3, 5.

So, in 2, 5, the distance is 5.4 this value, this value; 3, 5; 5.7, so we got this value, in between 5.4 and 5.7, minimum is 5.4. Now, 2, 3 versus 6, now we have to find out the distance between 2, 6 and 3, 6. So, 2, 6; where is 2, 6; 6, 2, this is 1.8, this value and 3, 6, 6, 3, see these 2 value, so in that minimum is 1.8. Now, the last point is 2, 3 versus 7, so what we have to do; we have to find the minimum distance is 2 and 7 and 3 and 7. So, between 2 and 7, minimum distance is 2.5, 3 and 7 minimum distance is 1.7, so out of this minimum is 1.7, now we are going to update this new distance.

(Refer Slide Time: 17:03)

Example for HAC

- Updated distance matrix for the cluster (2, 3)

	1,4	2,3	5	6	7
1,4	0.0				
2,3	4.0	0.0			
5	1.4	5.4	0.0		
6	5.8	1.8	7.2	0.0	
7	5.8	1.7	7.3	2.0	0.0

So, this was updated distance but you see that 2 and 3 is formed 1 cluster, previously 1 and 4 will formed a cluster, now in the next slides what we are going to do; among these new updated distance matrix which is the lowest value.

(Refer Slide Time: 17:17)

Example for HAC

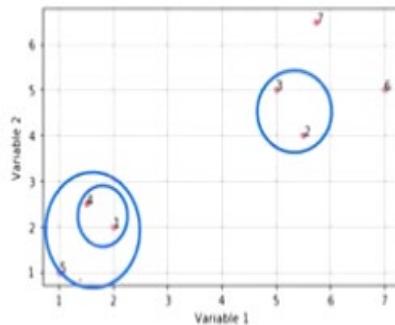
- Select minimum element to build next cluster formation-

	1,4	2,3	5	6	7
1,4	0.0				
2,3	4.0	0.0			
5	1.4	5.4	0.0		
6	5.8	1.8	7.2	0.0	
7	5.8	1.7	7.3	2.0	0.0

So, select the minimum element to build the next cluster formation, in that the minimum point is this 1.4, so what is going to happen; this object 5 is going to join with this cluster 1, 4.

(Refer Slide Time: 17:35)

Example for HAC



You see that the object 5 is going to form with cluster 1, 4.

(Refer Slide Time: 17:41)

Example for HAC

- Recalculate distance to update distance matrix

$$\begin{aligned} \text{- } \text{MIN}[\text{dist}((1,4),5), (2,3)] &= \text{MIN}(\text{dist}((1,4),(2,3)), (5,(2,3))) \\ &= \text{MIN}(4.0, 5.4) = 4.0 \end{aligned}$$

$$\text{- } \text{MIN}[\text{dist}((1,4),5), 6] = \text{MIN}(\text{dist}((1,4),6), (5,6)) = \text{MIN}(5.8, 7.2) = 5.8$$

$$\text{- } \text{MIN}[\text{dist}((1,4),5), 7] = \text{MIN}(\text{dist}((1,4),7), (5,7)) = \text{MIN}(5.8, 7.3) = 5.8$$

	1,4	2,3	5	6	7
1,4	0.0				
2,3	4.0	0.0			
5	1.4	5.4	0.0		
6	5.8	1.8	7.2	0.0	
7	5.8	1.7	7.3	2.0	0.0

Now, what we are going to do; you see that already there is; this is a new cluster, this is not new cluster, already 1, 4 is there, one cluster that again, the object 5 is added, so the distance between this cluster versus another cluster, in that there are 2 point; 2 object you see 2, 3. So, what we have to do; minimum distance between 1 and 4 and 2, 3 and 5 and 2, 3, so what is happening here; 1 and 4 is there, 2, 3 is there, the distance is 4, this value we got it.

The distance between 5, 2, 3 so this is 5.4, so this value we got it here, this value got it, the distant minimum is 4, similarly minimum distance between 1, 4, 5 versus 6, so what you have to do; 1, 4 versus 6 and 5 and 6, so 1, 4 versus 6, 1, 4 and 6, this is 5.8 right, so this value got here. So 5, 6; 7.2 we got this value here, the minimum is 5.8 now, there is a 7th object is there, let us see how far away the object 7. So, what you have to do; we have to find the minimum distance 1, 4 and 7; 1, 4 and 5, 7, so the distance between 1, 4 and 7 is 5.8 is this value and 5, 7 is our 7.3, so the minimum is 5.8.

(Refer Slide Time: 19:27)

Example for HAC

- Updated distance matrix for the cluster ((1,4), 5)

	1,4,5	2,3	6	7
1,4,5	0.0			
2,3	4.0	0.0		
6	5.8	1.8	0.0	
7	5.8	1.7	2.0	0.0

Again, we will update our distance matrix, so what happen now, you see that the 5 has entered into this cluster because already there is 1, 4 is there, so this is our updated distance matrix.

(Refer Slide Time: 19:43)

Example for HAC

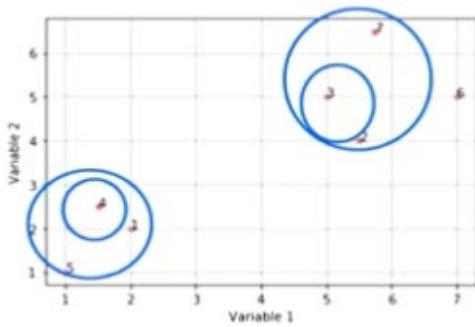
- Select minimum element to build next cluster formation-

	1,4,5	2,3	6	7
1,4,5	0.0			
2,3	4.0	0.0		
6	5.8	1.8	0.0	
7	5.8	1.7	2.0	0.0

The next step what we are going to do; in the updated distance matrix look at where there is a minimum distance is there, so the 1.7 is a minimum distance. So, what is going to do that the object 7 is going to add in to the cluster 2, 3, so now we will update this distance.

(Refer Slide Time: 20:05)

Example for HAC



What happen, see that this 7 is going to form in that cluster.

(Refer Slide Time: 20:09)

Example for HAC

- Recalculate distance to update distance matrix

	1,4,5	2,3	6	7
1,4,5	0.0			
2,3	4.0	0.0		
6	5.8	1.8	0.0	
7	5.8	1.7	2.0	0.0

$$\begin{aligned}
 & \text{- } \text{MIN}[\text{dist}((2,3), 7), (1,4,5)] = \text{MIN}(\text{dist}((2,3), (1,4,5)), (7, (1,4,5))) \\
 & \qquad\qquad\qquad = \text{MIN}(4.0, 5.8) = 4.0 \\
 & \text{- } \text{MIN}[\text{dist}((2,3), 7), 6] = \text{MIN}(\text{dist}((2,3), 6), (7, 6)) = \text{MIN}(1.8, 2.0) = 1.8
 \end{aligned}$$

Now, what happened, we are going to you see that already there is a cluster, in that 2 object is there because 7 also joined there, so the distance between this cluster versus 1, 4, 5, this is the another cluster. So, what we are going to do; so, 2, 3, 1, 4, 5 and 7, 1, 4, 5, so 2, 3 this 1; 2, 3, 1.45, yes the distance is 2, 3 1.45, this is; this 4 has come here. Second one; the distance between 7, 1, 4, 4, 5, so this one 5.8, so this distance has come here, out of this minimum is 4.

Now, this newly formed clustered versus 6, so here one point is 2, 3 versus 6 and 7, 6, so 2, 3 versus 6, what is a distance; 2, 3 versus 6 this is 1.8, so that distance is came here, so between 7

and 6, the distance is 2, so that distance has come here. Now, the minimum is 1.8, we cannot go next one because 7 is already gone in to that cluster.

(Refer Slide Time: 21:38)

Example for HAC

- Updated distance matrix for the cluster ((2,3), 7)

	1,4,5	2,3,7	6
1,4,5	0.0		
2,3,7	4.0	0.0	
6	5.8	1.8	0.0

Now, again we will update that now, in the updation you see that we have formed 2 clusters, in that 1, 4, 5 is one group, 2, 3, 7 is another group, this is updated distance matrix. This value 4 we got from here, this 1.8 we got from here.

(Refer Slide Time: 22:01)

Example for HAC

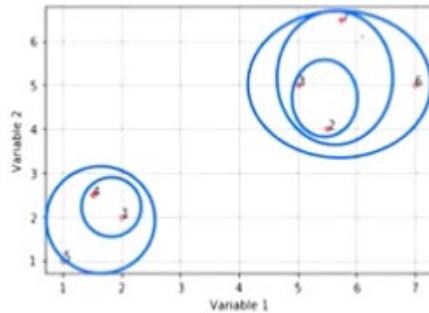
- Select minimum element to build next cluster formation-

	1,4,5	2,3,7	6
1,4,5	0.0		
2,3,7	4.0	0.0	
6	5.8	1.8	0.0

Now, in the next stage select minimum element to build the next cluster formation, so after this the minimum value is 1.8, this gives the minimum distance between 6 and the cluster 2, 3, 7, so what is going to do now; the 7 is going to join with this cluster where 2, 3, 7 is there.

(Refer Slide Time: 22:24)

Example for HAC



See that it is gone to 2, 3, 7.

(Refer Slide Time: 22:26)

Example for HAC

- Recalculate distance to update distance matrix

	1,4,5	2,3,7	6
1,4,5	0.0		
2,3,7	4.0	0.0	
6	5.8	1.8	0.0

$$\begin{aligned} - \text{MIN}[\text{dist}((2,3,7), 6), (1,4,5)] &= \text{MIN}(\text{dist}((2,3,7), (1,4,5)), (6, (1,4,5))) \\ &= \text{MIN}(4.0, 5.8) \\ &= 4.0 \end{aligned}$$

Now, recalculate the distance to update the distance matrix, so the distance between 2, 3, 7, 6 versus 1, 4, 5, so what you have to do the distance is 2, 3, 7 versus 1.45, you have to find out this distance, so 2, 3, 7, 1.45 the minimum distance is that one we brought it here. The next one 6, 1, 4.5, so 6, 1.45 that is a 5.8, so in that the minimum distance is 4.

(Refer Slide Time: 23:04)

Example for HAC

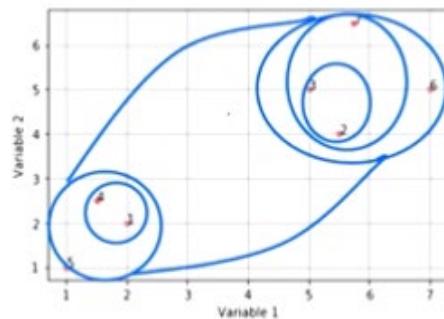
- Updated distance matrix for the cluster $((2,3,7), 6)$

	1,4,5	2,3,7,6
1,4,5	0.0	
2,3,7,6	4.0	0.0

This is our updated matrix, now what happened in that the minimum value is 4, so what happened this 2, 3, 7, 6 will join with 1, 4, 5.

(Refer Slide Time: 23:17)

Example for HAC



So that is nothing but the everything is joined together.

(Refer Slide Time: 23:22)

Python demo for HAC

```
In [1]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import scipy
from scipy.cluster.hierarchy import fcluster
from scipy.cluster.hierarchy import cophenet
from scipy.spatial.distance import pdist

In [2]: data = pd.read_excel("hierarchical_clustering.xlsx")
data

Out[2]:
```

	Variable 1	Variable 2
0	2.00	2.0
1	5.50	4.0
2	5.00	5.0
3	1.50	2.5
4	1.00	1.0
5	7.00	5.0
6	5.75	6.5

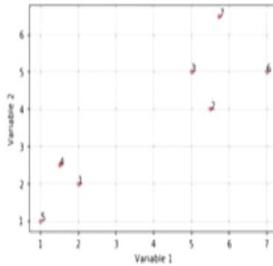
Now, let us see how to do this agglomerative hierarchical clustering method with the help of python, so we have imported the data; import numpy as np, import pandas as pd, import matplotlib.pyplot as plt, import scipy. from scipy.cluster.hierarchy import fcluster, from scipy.cluster.hierarchy import cophenet, from scipy.spatial.distance import pdist.

(Refer Slide Time: 24:07)

Python demo for HAC

```
In [3]: x = data['Variable 1']
y = data['Variable 2']
n = range(1,8)

fig, ax = plt.subplots()
ax.scatter(x, y, marker='^', c='red', alpha=0.5)
plt.grid()
plt.xlabel("Variable 1")
plt.ylabel("Variable 2")
for i, txt in enumerate(n):
    ax.annotate(txt, (x[i], y[i]))
```

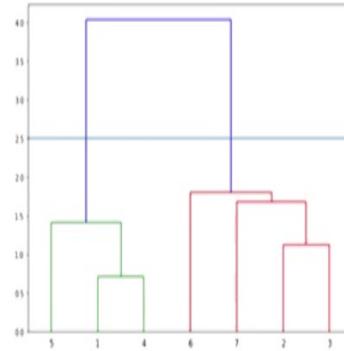


So, I have data in hierarchical clustering, so what happened this was our data set, for this data set, we have plotted the 2 dimensional picture, so in that all the objects are displayed.

(Refer Slide Time: 24:14)

Python demo for HAC

```
In [4]: from scipy.cluster.hierarchy import dendrogram, linkage  
linked = linkage(data, 'single')  
labellist = range(1, 8)  
plt.figure(figsize=(10, 7))  
dendrogram(linked,  
           orientation='top',  
           labels=labellist,  
           distance_sort='descending',  
           show_leaf_counts=True)  
plt.axhline(y=2.5)  
plt.show()
```



Now, when you run this comment that is from `scipy.cluster.hierarchy import dendrogram, linkage`, you will get a this kind of pictures, So, what is happening you see that this is the way 1 and 4 are joining together this stage, after that along with 1 and 4, the 5 is joined, initially 2 and 4; 2 and 3 is a joined, so along with 2 and 3, 7 is joined, after sometime along with the 7, 2, 3, 6 also joined.

At the end, see that the blue line which says all are forming one clustering, this picture shows the dendrogram, so for that from `scipy.cluster.hierarchy import dendrogram, linkage`, so linked equal to `linkage(data, 'single')`, we are going to have single linkage, the label is this, range 1 to 8 that is a figure size. So, when you run that you are getting the dendrogram.

(Refer Slide Time: 25:17)

Python demo for HAC

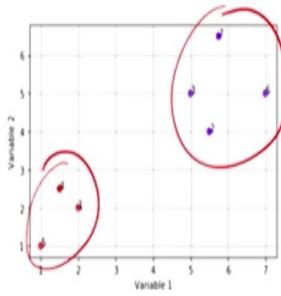
```
In [5]: import sklearn  
from sklearn.cluster import AgglomerativeClustering  
  
k=2  
Hclustering = AgglomerativeClustering(n_clusters = k, affinity = 'euclidean', linkage = 'single')  
Hclustering.fit(data)  
  
Out[5]: AgglomerativeClustering(affinity='euclidean', compute_full_tree='auto',  
connectivity=None, distance_threshold=None,  
linkage='single', memory=None, n_clusters=2,  
pooling_func='deprecated')  
  
In [6]: Hclustering.fit_predict(data)  
Out[6]: array([1, 0, 0, 1, 1, 0], dtype=int64)  
  
In [7]: print(Hclustering.labels_)  
[1 0 0 1 1 0]
```

Now, here what it shows that, here we have entered k equal to 2, you look at this value from sklearn; from sklearn dot cluster import AgglomerativeClustering, when you put k equal to 2, we are going to say Euclidean distance and single linkage, so when you; k equal to 2, so the cluster name is into 2 category; one is 0 is one group, 1 is another group, so this was the labels; 1, 0.

(Refer Slide Time: 25:46)

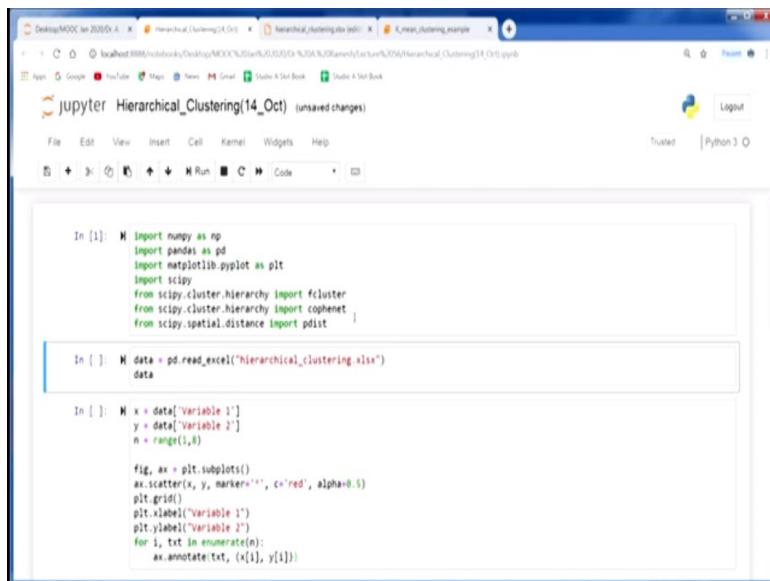
Python demo for HAC

```
In [8]: x = data['Variable 1']  
y = data['Variable 2']  
n = range(1,8)  
  
fig, ax = plt.subplots()  
ax.scatter(x, y, c=Hclustering.labels_, cmap='rainbow')  
plt.grid()  
plt.xlabel("Variable 1")  
plt.ylabel("Variable 2")  
for i, txt in enumerate(n):  
    ax.annotate(txt, (x[i], y[i]))
```



So, when you run this, you see that there are 2 clustering, so this is forms 1 cluster, this is form another cluster, suppose if you write k equal to 3 here, you may get with 3 clusters now, I am going to the python demo for doing this agglomerative hierarchical clustering.

(Refer Slide Time: 26:11)

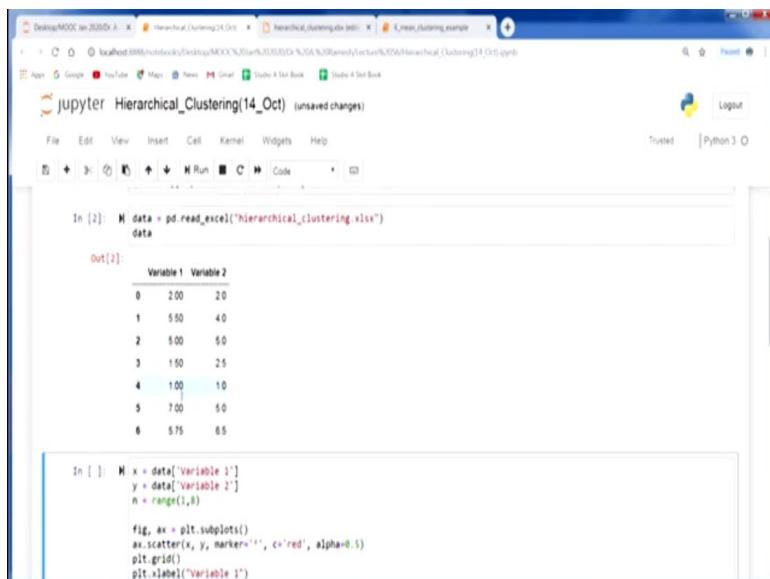


The screenshot shows a Jupyter Notebook interface with two code cells. The first cell contains imports for numpy, pandas, matplotlib.pyplot, and various clustering functions from scipy. The second cell reads data from an Excel file named 'hierarchical_clustering.xlsx' into a pandas DataFrame named 'data'. The code then extracts 'Variable 1' and 'Variable 2' from the data and creates a scatter plot.

```
In [1]: import numpy as np  
import pandas as pd  
import matplotlib.pyplot as plt  
import scipy  
from scipy.cluster.hierarchy import fcluster  
from scipy.cluster.hierarchy import cophenet  
from scipy.spatial.distance import pdist  
  
In [2]: data = pd.read_excel("hierarchical_clustering.xlsx")  
data
```

Now, I am going to show how to do agglomerative hierarchical clustering in python, so import this necessary library; I am running this, I have stored the data in the file name hierarchical_clustering.

(Refer Slide Time: 26:26)



The screenshot shows a Jupyter Notebook interface with two code cells. The first cell reads data from 'hierarchical_clustering.xlsx' into a pandas DataFrame named 'data'. The second cell displays the data as a table:

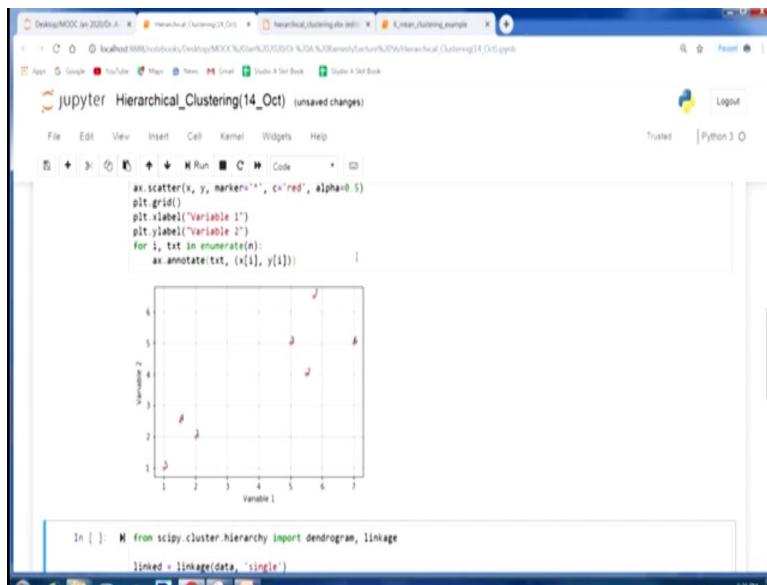
	Variable 1	Variable 2
0	200	20
1	550	40
2	500	60
3	150	25
4	100	10
5	700	50
6	575	65

The second cell contains the same code as the previous screenshot, reading the data and creating a scatter plot.

```
In [2]: data = pd.read_excel("hierarchical_clustering.xlsx")  
data  
  
Out[2]:  
Variable 1 Variable 2  
0 200 20  
1 550 40  
2 500 60  
3 150 25  
4 100 10  
5 700 50  
6 575 65  
  
In [3]: x = data['Variable 1']  
y = data['Variable 2']  
n = range(1,8)  
  
fig, ax = plt.subplots()  
ax.scatter(x, y, marker='*', c='red', alpha=0.5)  
plt.grid()  
plt.xlabel("Variable 1")
```

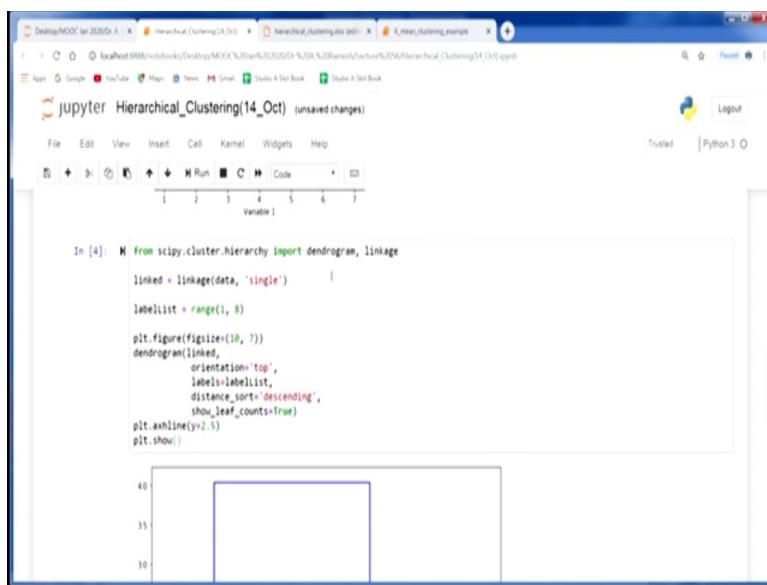
So, this is our data, for this I am going to do this hierarchical clustering.

(Refer Slide Time: 26:33)



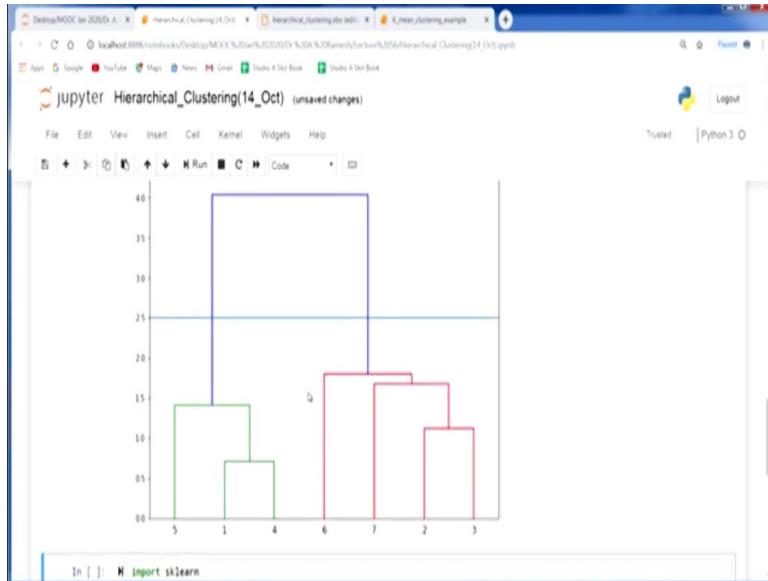
So, first I am going to show the scatterplot, in that it is showing all the objects, while looking at the object itself, you see that if you go for 2 clustering that will be good.

(Refer Slide Time: 26:46)



So, what we are going to do; we are going to do the single linkage, here I am going to show the figure size here also, so this picture you know we are going to get the dendrogram.

(Refer Slide Time: 26:56)



So, this is the dendrogram, so 2 and 3 is forming one cluster, out of that this 7 is joining, out of that 6 is joining. Here at level 1, it is 1 and 4 is joining, then 5 is joining there at the end, it is going to; all are going to be in the same cluster.

(Refer Slide Time: 27:12)

```
In [5]: H = import sklearn
         from sklearn.cluster import AgglomerativeClustering

         k=2
         Hclustering = AgglomerativeClustering(n_clusters = k, affinity = 'euclidean', linkage = 'single')
         Hclustering.fit(data)

Out[5]: AgglomerativeClustering(affinity='euclidean', compute_full_tree='auto',
                                 connectivity=None, linkage='single', memory=None, n_clusters=2,
                                 pooling_func='deprecated')

In [6]: Hclustering.fit_predict(data)

In [7]: print(Hclustering.labels_)

In [8]: x = data['Variable 1']
        y = data['Variable 2']
        n = range(1,8)
```

Now, from sklearn dot cluster import AgglomerativeClustering, suppose we will start with 2, so 2, so run this, let us see hierarchical clustering, how we are doing, see that there are 2 clustering, it is one is named as 1, another one is 0.

(Refer Slide Time: 27:38)

```

Out[5]: AgglomerativeClustering(affinity='euclidean', compute_full_tree='auto',
                                 connectivity=None, linkage='single', memory=None, n_clusters=2,
                                 pooling_func='deprecated')

In [6]: M = clustering.fit_predict(data)

Out[6]: array([1, 0, 0, 1, 1, 0], dtype=int64)

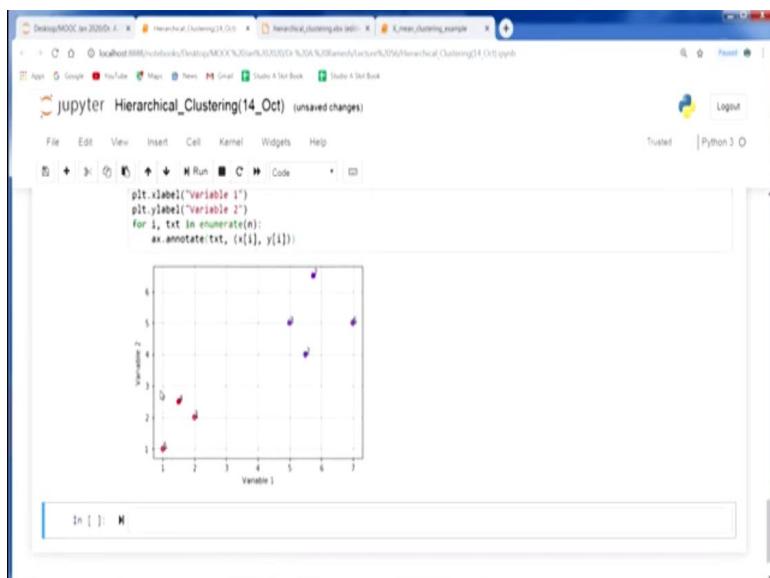
In [7]: print(Mclustering.labels_)

In [8]: M x = data['Variable 1']
y = data['Variable 2']
n = range(1,8)
fig, ax = plt.subplots()
ax.scatter(x, y, c=Mclustering.labels_, cmap='rainbow')
plt.grid()
plt.xlabel('Variable 1')
plt.ylabel('Variable 2')
for i, txt in enumerate(n):
    ax.annotate(txt, (x[i], y[i]))

```

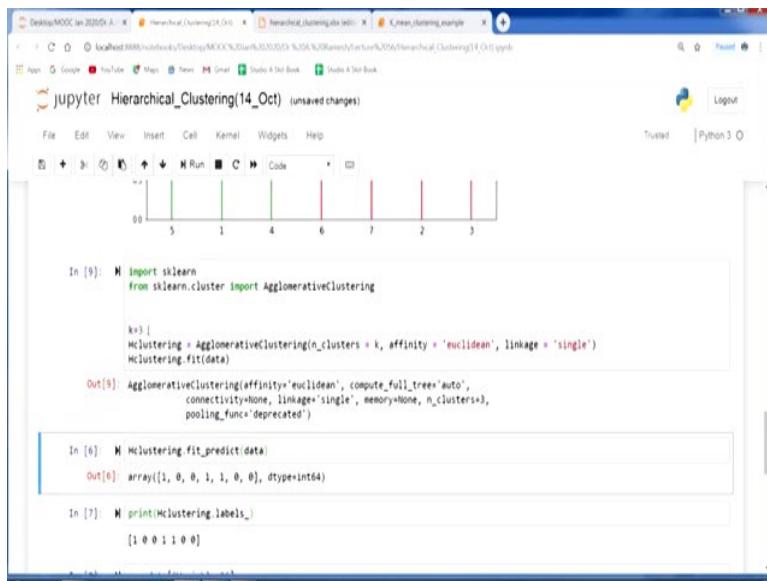
So, let us see what are the labels so, this is labels; 1, 0, 0, 1, 1, 0, 0.

(Refer Slide Time: 27:48)



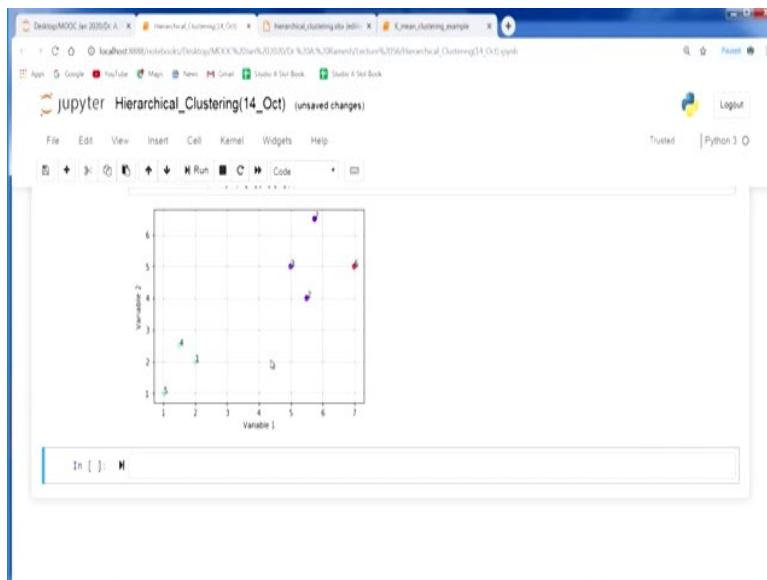
So, now if you run this, what you are getting; now there are 2 clusters, so the red colour point, say 1 cluster, the purple is another cluster. Suppose, if we go for k equal to 3, let us see what kind of answer we are getting.

(Refer Slide Time: 28:04)



Suppose, if we go for k equal to 3, again you run it, you see that the level is 0, 1, 2, again you run it, level is see that.

(Refer Slide Time: 28:18)



Now, if you run this, you see that there are 3 cluster; green, purple and red because red is only one cluster, there is only one element, so the optimal number of cluster for this kind of data set is k equal to 2 that is a purpose we can visualise how the cluster is formed and quality of cluster also. In this lecture, what I have done, I have started agglomerative hierarchical algorithm with the help of a numerical example.

In that example, I have first I found the distance matrix, after finding the distance matrix, I formed a cluster wherever there is a minimum value is there, I connected that 2 objects, then I have updated the distance matrix, again I have gone to where there is a minimum point is there, so that point and that objects are clubbed together. At the end, for the same data set, I have explained how to do python programming.

And I also shown how the result is appearing, so here the number of cluster I have initially started with k equal to 2, then again I changed k equal to 3 but when I changed k equal to 3, I get some other result that is not looking good, so I kept only k equal to 2 is the right number of clusters, so optimal number of clusters.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology - Roorkee

Lecture – 57
Classification and Regression Trees(CART) - I

In our previous lecture, we studied about different cluster and techniques, in this class we will start a new topic that is a classification and regression trees, shortly this is called CART models.

(Refer Slide Time: 00:41)

Agenda

- Introduction to Classification and Regression Trees
- Attribute selection measures – Introduction

The agenda for this lecture is introduction to classification, regression trees, attribute selection measures and introduction. There are different measures for selecting attributes; attributes means variables that we will study about different attribute selection measures in this class.

(Refer Slide Time: 00:55)

Introduction

- Classification is one form of data analysis that can be used to extract models describing important data classes or to predict future data trends
- Classification predicts categorical (discrete, unordered) labels whereas Regression analysis is a statistical methodology that is most often used for numeric (continuous) prediction
- For example, we can build a classification model to categorize bank loan applications as either safe or risky
- Regression model is used to predict expenditures in dollars of potential customers on computer equipment given their income and occupation

The introduction about this topic, CART model; classification is one form of data analysis that can be used to extract models describing important data classes or to predict future data trends. Classification predicts categorical that is discrete, unordered labels, whereas regression analysis is a statistical methodology that is most often used for numeric prediction. There is a difference between classification techniques and regressions.

In a classification techniques; the dependent variable is categorical variable most of the time but in regression analysis, most of the time the regression analysis, the continuous variable is the dependent variable for regression analysis. For example, we can build a classification model to categorize the bank loan applications as either safe or risky, see this is categorical. The regression model is used to protect expenditures in dollars of potential customers and computed equipment given their income and occupations. Most of the time the regression analysis used to predict a continuous variable but the classification analysis is used to predict the categorical variable.

(Refer Slide Time: 02:16)

Problem Description for Illustration

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Han, J., Pei, J. and Kamber, M., 2011. Data mining: concepts and techniques. Elsevier.

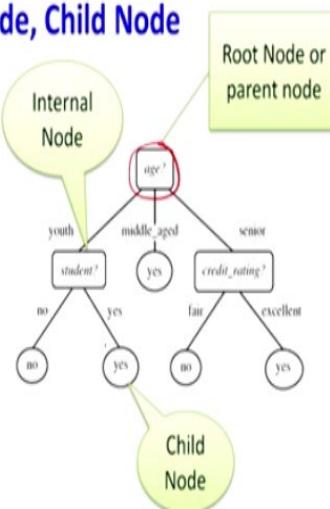
We are going to take one problem, this problem is taken from this book, Han, Pei and Kamber, data mining; book title is data mining concept and techniques. The problem says there are 1, 2, 3, 4, 5 columns, in the 5 columns there is age is there, income is there, student, credit rating, this dependent variable is buys computer; buys underscore computer, so this is a database, one portion of data base is shown.

So, the dependent variable is buys computer, there are 4 independent variables like age, income, student and credit rating. So by taking this example, we are going to explain how to use CART model, in coming lecture also we will use this data.

(Refer Slide Time: 03:04)

Root Node, Internal Node, Child Node

- A decision tree uses a tree structure to represent a number of possible *decision paths* and an outcome for each path
- A decision tree consists of root node, internal node and leaf node
- The topmost node in a tree is the **root node** or **parent node**
- It represents entire sample population
- **Internal node** (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test
- **Leaf node** (or **terminal node** or **child node**) holds a class label
- It can not be further split



Now, let us understand certain terminology in the CART model, for example root node, internal node and child node, when you look at this picture, there is age; age there are 3 levels; youth, middle aged, senior, then it is a student there are 2 levels; no or yes. Credit rating; there are 2 levels; fair and excellent. So, a decision tree uses a tree structure to represent a number of possible decision paths and an outcome for each path.

The decision tree consist of root node, internal node and leaf node, you look at this the first one the age because the whole problem we are going to start with a variable age, so this age is called root node or parent node that is in the rectangular box. A decision tree consist of root node, internal node and leaf node, the top most node in a tree is called root node or parent node, this one for example age, this is a root node.

It represents entire sample population, the next term is internal node, for example here student is the internal node or non-leaf node, denotes a test on an attribute, each branch represents outcome of the test. The next node is leaf node or child node, see this yes or no that is which is in the elliptical shape that is called leaf node, it cannot be further split.

(Refer Slide Time: 04:46)

Decision Tree Introduction

- A decision tree for the concept *buys_computer*, indicating whether a customer at All Electronics is likely to purchase a computer
- Each internal (non-leaf) node represents a test on an attribute
- Each leaf node represents a class (either *buys_computer* = yes or *buys computer* = no).

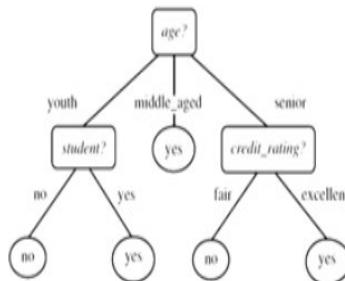


Figure 1.1 : Decision Tree

A decision tree for the concept buys computer is a variable indicating whether a customer at all electronics that is a database, where it was taken, a customer at all electronics is likely to purchase a computer, each internal node represents a test on attribute, each leaf node represents a

class, a class means either a person buys the computer yes or no, actually this yes or no is nothing but this column that we will go to into the child node.

(Refer Slide Time: 05:21)

CART Introduction

- CART comes under supervised learning technique
- CART adopt a greedy(i.e., non backtracking) approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner
- It is very interpretable model

Now, the CART comes under supervised learning techniques, we know that the machine learning techniques are classified into 2 categories; one is supervised, another one is unsupervised. What is the meaning of supervised learning is that there is a label; label in the sense we know in advance what is going to be independent variable, what is going to be dependent variable, in this problem also, it is supervised learning.

Because we know in advance what is the buys underscore computer is going to be our dependent variable, then CART adopts a greedy that is a non-backtracking approach in which decision trees are constructed in a top down recursive divide and conquer manner, it is very interpretable model. A person who was not having any statistical analysis also can easily interpret the CART model.

(Refer Slide Time: 06:14)

Decision Tree Algorithm

Input:

- Data partition, D, which is a set of training tuples and their associated class labels;
- Attribute list, the set of candidate attributes;
- Attribute selection method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting attribute and, possibly, either a split point or splitting subset.

RID	age	income	student	credit rating	Class buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle aged	medium	no	excellent	yes
13	middle aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Output: A decision tree

Now, I will explain the decision tree algorithm, what are the inputs data partition D, this whole data set is a set of training tuples and their associated class labels, this is the class labels. Attribute list is the set of candidate attribute, for example in this problem these are the attributes; age, income, student, credit rating. Now, what you have to do before starting the problem, out of these 4 variables, we have to decide from which variable we have to start for classification.

So, for that we need a attribute selection method, so attribute selection method a procedure to determine the splitting criterion that best partitions the data tuples into individual classes, this criterion consist of a splitting the attribute and possibly, either a split point on splitting subset. So, output of this model will be the decision tree.

(Refer Slide Time: 07:16)

Decision Tree Algorithm

- The algorithm is called with three parameters: D, attribute list, and Attribute selection method
- D is defined as a data partition. Initially, it is the complete set of training tuples and their associated class labels
- The parameter attribute list is a list of attributes or independent variables which are describing the tuples
- Attribute selection method specifies a heuristic procedure for selecting the attribute that “best” discriminates the given tuples according to class

Decision tree algorithm; the algorithm is called with the 3 parameters; one is D that is a data set, then is attribute list independent variable and attribute selection methods. D is defined as data partition; initially it is the complete set of training tuples and their associated class labels. In our problem, this whole table represents the D, initially what is that it covers all the independent variables and dependent variables.

The parameter attribute list is a list of attributes or independent variables which are describing the tuples. Here the parameter attributes; attributes nothing but all the independent variables, so attribute selection method specify a heuristic procedure for selecting the attribute that best discriminates the given tuples according to class.

(Refer Slide Time: 08:12)

Decision Tree Algorithm



- This procedure employs an attribute selection measure, such as information gain, gain ratio or the Gini index.
- Whether the tree is strictly binary is generally driven by the attribute selection measure
- Some attribute selection measures, such as the Gini index, enforce the resulting tree to be binary. Others, like information gain, do not, therein allowing multiway splits (i.e., two or more branches to be grown from a node).

This procedure employs an attribute selection measures such as information gain, that is a one method for selecting the attribute, second one method is called gain ratio, third method is called Gini index, whether the tree is strictly binary is generally driven by the attribute selection measures. For example, we can go for binary selection suppose, this is one variable, sometime we can go for more than 2 classifications also.

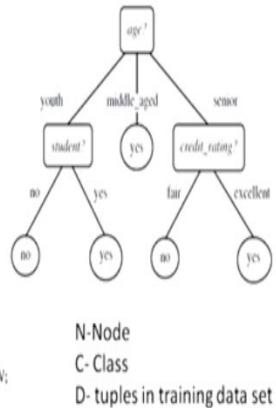
For example, if you use Gini method that will cover in coming class that always you need to go for binary selection, some attribute selection measures such as Gini index enforce the resulting tree to be binary, others like information gain do not, therein allowing multiway splits. So, if you follow Gini index, there is only binary split, if you follow other than Gini index for example, information gain, you can have more than 2 split also.

(Refer Slide Time: 09:11)

Decision Tree Method

Method:

- (1) create a node N ;
- (2) if tuples in D are all of the same class, C then
- (3) return N as a leaf node labeled with the class C ;
- (4) if $attribute_list$ is empty then
- (5) return N as a leaf node labeled with the majority class in D ; // majority voting
- (6) apply Attribute.selection.method($D, attribute_list$) to find the "best" splitting.criterion;
- (7) label node N with splitting.criterion;
- (8) if splitting.attribute is discrete-valued and
 multiway splits allowed then // not restricted to binary trees
- (9) $attribute_list \leftarrow attribute_list - splitting.attribute$, // remove splitting.attribute
- (10) for each outcome j of splitting.criterion
 // partition the tuples and grow subtrees for each partition
- (11) let D_j be the set of data tuples in D satisfying outcome j ; // a partition
- (12) if D_j is empty then
- (13) attach a leaf labeled with the majority class in D to node N ;
- (14) else attach the node returned by Generate.decision.tree($D_j, attribute_list$) to node N ;
- endfor
- (15) return N ;



N-Node
 C- Class
 D- tuples in training data set

Decision tree method; I am going to explain different steps in the decision tree method, there are 15 steps in coming slides, I will explained in each steps. So, first we will start the overview of all 15 steps. Create a node N , if tuples in D all are of the same class C , then return N as the leaf node labelled with the class C , if attribute list is empty then return N as a leaf node labelled with majority class in D by using the concept called majority voting.

Then apply attribute selection method D to find the best splitting criterion, label node N with the splitting criterion, if splitting attribute is discrete valued and multiway splits is allowed, then attribute list, there are different attribute list that we can choose for example, splitting attribute is one method. The step 8 is if splitting attribute is discrete valued and multiway split is allowed, then the attribute list which already have occurred that has to be removed from the our D .

For each outcome j for splitting criterion, so what is a splitting criterion is partition of the tuples and grow sub tree for each partition. Let D_j be the set of data tuples in D satisfying the outcome j , if D_j is empty then attach a leaf labelled with the majority class in D to node N else attach node returned by generate decision tree to node N , then N for return N . So, I am going to explain that each steps in detail in coming slides.

(Refer Slide Time: 11:05)

Decision Tree Method step 1 to 6

- The tree starts as a single node, N, representing the training tuples in D (step 1).
- If the tuples in D are all of the same class, then node N becomes a leaf and is labelled with that class (steps 2 and 3)
- Steps 4 and 5 are terminating conditions
- Otherwise, the algorithm calls Attribute selection method to determine the splitting criterion
- The splitting criterion (like Gini) tells us which attribute to test at node N by determining the "best" way to separate or partition the tuples in D into individual classes (step 6)

Method:

- (1) create a node N_c
- (2) if tuples in D_c are all of the same class, C then
 (3) return N_c as a leaf node labelled with the class C ;
- (4) if attribute list is empty then
 (5) return N_c as a leaf node labelled with the majority class in D_c ; // majority voting
- (6) apply Attribute selection method(D_c , attribute list) to find the "best" splitting criterion;
- (7) label node N_c with splitting criterion;
- (8) if splitting attribute is discrete-valued and
 multway splits allowed then // not restricted to binary trees
- (9) attribute list \leftarrow attribute list - splitting attribute; // remove splitting attribute
- (10) for each outcome j of splitting criterion
 // partition the tuples and grow subtrees for each partition
- (11) let $D_{c,j}$ be the set of data tuples in D_c satisfying outcome j ; // a partition
- (12) if $D_{c,j}$ is empty then
- (13) attach a leaf labelled with the majority class in D_c to node N_c ;
- (14) else attach the node returned by Generate decision tree($D_{c,j}$, attribute list) to node N_c ;
- endfor
- (15) return N_c

The tree starts as a single node N representing training tuples in D that was our step 1, if the tuples in D are all of the same class, then N becomes a leaf and is labelled with that class that is step 2 and 3. If the tuples in D, all of the same class C, then return N as a leaf node labelled with the class C, the meaning is, suppose the age is taken as variable, if there are 3 split; one is youth, middle-aged and senior.

For example, the middle-aged with respect to our dependent variable all are answered yes, so if it all are answered yes, we need not go for further classification, then the age attribute has to be dropped from our model, then we have continue with the remaining attributes like student, credit rating and income. So, step 4 and 5 are terminating conditions, if attribute list is empty that means, you have to go for each attributes otherwise, the algorithm calls attribute selection method to determine splitting criterion.

Suppose, only 1 attribute is there, if there are remaining attributes, to choose that attribute, you have to use attribute selection method to determine splitting criterion. The splitting criterion like Gini tells us which attribute to test at node N by determining the best way to separate or partition the tuples in D into individual classes.

(Refer Slide Time: 12:44)

Decision Tree Method - Step 7 - 11

- The splitting criterion indicates the splitting attribute and may also indicate either a split-point or a splitting subset
- The splitting criterion is determined so that, ideally, the resulting partitions at each branch are as "pure" as possible. A partition is pure if all of the tuples in it belong to the same class.
- The node N is labelled with the splitting criterion, which serves as a test at the node (step 7).
- A branch is grown from node N for each of the outcomes of the splitting criterion.
- The tuples in D are partitioned accordingly (steps 10 to 11)

Method:

```

(1) create a node N;
(2) if tuples in D are all of the same class, C then
    (3) return N as a leaf node labeled with the class C;
(4) if attribute list is empty then
    (5) return N as a leaf node labeled with the majority class in D; // majority voting
(6) apply Attribute selection method(D, attribute list) to find the "best" splitting criterion;
(7) label node N with splitting criterion;
(8) if splitting attribute is discrete-valued and
        multivalue splits allowed then // not restricted to binary trees
            attribute list -- attribute list - splitting attribute; // remove splitting attribute
(9) for each outcome j of splitting criterion
    // partition the tuples and grow subtrees for each outcome j; // a partition
(10) let  $D_j$  be the set of data tuples in D satisfying outcome  $j$ ; // a partition
(11) if  $D_j$  is empty then
    (12) attach a leaf labeled with the majority class in D to node N;
(13) else attach the node returned by Generate decision tree( $D_j$ , attribute list) to node N;
    endfor
(14) return N;

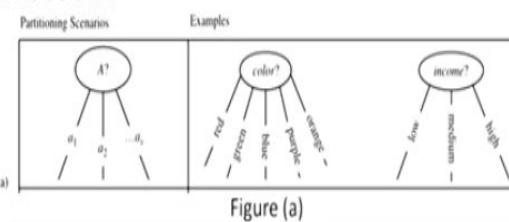
```

The splitting criterion indicates the splitting attributes and may also indicate either a split pointer or splitting subset; I will explain, what is the meaning of split point and splitting subset in next slide. The splitting criterion is determined, so that ideally the resulting partitions at each branch are as a pure as possible. A partition pure if all of the tuples in it belongs to the same class, the node N is labelled with the splitting criterion which serve as a test at the node that is a step 7. A branch is grown from node N for each of the outcomes for splitting criterion, the tuples in D are partitions accordingly that is our step in 11.

(Refer Slide Time: 13:30)

Three possibilities for partitioning tuples based on the splitting criterion

- There are three possible scenarios, as illustrated in Figure (a), (b) and (c).
- Let A be the splitting attribute. A has 'v' distinct values, {a₁, a₂, ..., a_v}, based on the training data
- If A is discrete-valued in figure (a), then one branch is grown for each known value of A.



So, 3 possibilities for partitioning tuples based on the splitting criterion, there are 3 possible scenarios as illustrated in figure a, b and c. Let A be the splitting attributes, A has v distinct

values a_1 ; see a_2, a_3 and a_v based on training data. If A is discrete valued in figure a, then one branch is grown for each known value of A . See for example colour; may be red, green, blue, purple, orange, if it is income, there are 3 split; low, medium, high.

(Refer Slide Time: 14:13)

Three possibilities for partitioning tuples based on the splitting criterion

- If A is continuous-valued in figure (b), then two branches are grown, corresponding to $A \leq \text{split point}$ and $A > \text{split point}$.
- Where split point is the split-point returned by Attribute selection method as part of the splitting criterion.



Figure (b)

If A is a continuous valued in figure b, then 2 branches are grown corresponding to A less than or equal to split point and A greater than or equal to split point. So, what will happen; the A less than or equal to split point is the one split, A greater than split point is another branch, where the split point is the split point returned by attribute selection method as part of the splitting criterion.

For example, income is there, we can group that into 2 categories, those who have incomes are below 42,000, those who have incomes are above 42,000; this 42,000 generally is nothing but the average value.

(Refer Slide Time: 14:54)

Three possibilities for partitioning tuples based on the splitting criterion

- If A is discrete-valued and a binary tree must be produced, then the test is of the form $A \in S_A$, where S_A is the splitting subset for A.

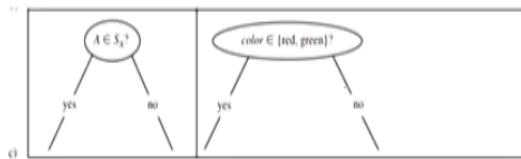


Figure (c)

If A is a discrete valued and binary tree must be produced, then the test is of the form A belongs to S_A , where S_A is the splitting subset of A, so is it that A belongs to S_A , if it is yes is one group, no if it is another group. In the; if A, for example colour it may be red or green, then that time also, it should be yes or no.

(Refer Slide Time: 15:18)

Decision Tree Method – termination condition

- The algorithm uses the same process recursively to form a decision tree for the tuples at each resulting partition, D_j , of D (step 14).
- The recursive partitioning stops only when anyone of the following terminating conditions is true:
 1. All of the tuples in partition D (represented at node N) belong to the same class (steps 2 and 3), or

Then, we will go for termination condition; the algorithm uses the same process recursively to form a decision tree for the tuples at each resulting partition D_j of D, what is the recursion means; if one attribute is over that is repeated for the second attribute and third attribute up to all the attributes are exhausted. The recursive partitioning stops only when any one of the following

terminating condition is true. The first condition is all of the tuples in partition D representing at node N belong to the same class that was our same step 2 and 3.

(Refer Slide Time: 15:58)

Decision Tree Method – termination condition

2. There are no remaining attributes on which the tuples may be further partitioned (step4).
 - In this case, majority voting is employed(step 5).
 - This involves converting node N into a leaf and labelling it with the most common class in D.
 - Alternatively, the class distribution of the node tuples may be stored.
3. There are no tuples for a given branch, that is, a partition Dj is empty (step 12).
 - In this case, a leaf is created with the majority class in D (step 13).
 - The resulting decision tree is returned (step 15).

Or there are no remaining attribute on which the tuples maybe further partitioned that is in step 4, in this case majority voting is employed, this involves converting a node into a leaf and labelling it with the most common class in D, alternatively the class distribution of the node tuples may be stored. The third condition is there are no tuples for a given branch that is a partition Dj is empty that is explained in step 12. In this case, a leaf is created with the majority class in D that is your step 13, the resulting decision tree is returned that is our step 15.

(Refer Slide Time: 16:42)

Attribute Selection Measures

- Attribute selection measures are also known as splitting rules because they determine how the tuples at a given node are to be split
- It is a heuristic approach for selecting the splitting criterion that “best” separates a given data partition, D, of class-labeled training tuples into individual classes
- The attribute selection measure provides a ranking for each attribute describing the given training tuples
- The attribute having the best score for the measure is chosen as the splitting attribute for the given tuples

Now, the second part of the selection is different attribute selection measures; attribute selection measures are also known as splitting rules because they determine how the tuples at a given node to be split, it is a heuristic approach for selecting the splitting criterion that best separates a given data partition D of class labelled training tuples into individual classes. The attribute selection measures provide a ranking for each attributes describing the given training tuples. The attributes having the best score for the measure is chosen as the splitting attribute for the given tuples.

If the splitting attribute is continuous valued or if we are restricted to binary trees, then respectively, either a split point or split subset must also be determined as part of the splitting criterion. There are 3 popular attribute selection measures; one is information gain, gain ratio, Gini index. In this class, I am going to explain the theory about this 3 attribute measures, in coming classes by using the same examples which I have discussed, I am going to find out the value of information gain.

I am going to explain the selection procedures using the criteria information gain, gain ratio and Gini index. So, in this lecture this we are going to see the theoretical point of all these 3 selection methods. So, CART algorithm uses Gini index measures for attribute selection.

(Refer Slide Time: 18:30)

Attribute Selection Measures

The notation used herein is as follows. Let D , the data partition, be a training set of class-labeled tuples. Suppose the class label attribute has m distinct values defining m distinct classes, C_i (for $i = 1, \dots, m$). Let $C_{i,D}$ be the set of tuples of class C_i in D . Let $|D|$ and $|C_{i,D}|$ denote the number of tuples in D and $C_{i,D}$, respectively.

ID	age	income	student	credit rating	Class	buys computer
1	youth	high	no	fair	no	no
2	youth	high	no	excellent	no	no
3	middle aged	high	no	fair	yes	yes
4	senior	medium	no	fair	yes	yes
5	senior	low	yes	fair	yes	yes
6	senior	low	yes	excellent	no	no
7	middle aged	low	yes	excellent	yes	yes
8	youth	medium	no	fair	no	no
9	youth	low	yes	fair	yes	yes
10	senior	medium	yes	fair	yes	yes
11	youth	medium	yes	excellent	yes	yes
12	middle aged	medium	no	excellent	yes	yes
13	middle aged	high	yes	fair	yes	yes
14	senior	medium	no	excellent	no	no

$m = 2$

Attribute selection measures let us find out certain notations, the notation used herein is as follows. Let D , the data partitions, be a training set of class labelled tuples, for example this

dataset suppose, the class label attribute as m distinct values, here m is there are; this is a class 1, here the value of m is 2 because yes is 1 category, no is another category, distinct class in Ci. The Ci is it is 1 to m, it may be 1 and another may be it is 2.

Let CiD be the set of tuples of class CiD for example, what is the CiD means, for example if it is a high for this income variable, in high how many no is there; 1, 2, high, it is a 2, for income one level is called high for that it is no, no, so that is our CiD; CiD, set of tuples of class Ci in D. So, this modulus of D represents that the 14 CiD represents how many number of values if it is high and what is no. If it is for example, if you say low, this variable; income variable low, how many yes is there; 1, low, 2, low, 3, so 3, the modulus of CiD is 3, so this values I have explained in coming lectures with an example.

(Refer Slide Time: 20:11)

Information Gain

- This measure studied the value or "information content" of messages
- The attribute with the highest information gain is chosen as the splitting attribute for node
- This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or "impurity" in these partitions
- This approach minimizes the expected number of tests needed to classify a given tuple

Then, we will go to the first criteria for selecting the attributes information gain, this measure studied the value or information content of messages, the attribute with the highest information gain is chosen as the splitting attribute for node, this attribute minimises the information needed to classify the tuples in the resulting partitions and reflect the least randomness or impurity in these partitions.

(Refer Slide Time: 20:49)

Information Gain-Entropy Measure

- The expected information needed to classify a tuple in D is given by

$$Info(D) = - \sum_{i=1}^m p_i \log_2(p_i)$$

- Where p_i is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_{i,D}|/|D|$.
- A log function to the base 2 is used, because the information is encoded in bits
- Info(D) (or Entropy of D) is just the **average amount of information needed** to identify the class label of a tuple in D

ID	age	income	student	credit rating	Class: buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle aged	medium	no	excellent	yes
13	middle aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

$$Info(D) = -\frac{9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits.}$$

So, this approach minimises the expected number of test needed to classify a given tuple, so information gain that is nothing but entropy measure, the expected information needed to classify a tuple in D is given by $Info(D) = -\sum_{i=1}^m p_i \log_2(p_i)$ to the base 2, where p_i is a probability that an arbitrary tuple in D belongs to class C_i and is estimated by modulus of $(C_i D)$ divided by modulus of (D) .

For example, in this for this dataset, what is a p_i ; p_i is the number of yes, how many number of yes is there? 1, 2, 3, 4, 5, 6, 7, 8, 9, so it is 9, that 9 is nothing but your $C_i D$, modulus of (D) is total 14 that is for level 1. For the level 2, how many no is there? 1, 2, 3, 4, 5, so 5 divided by 14 log 5 divided by 14 to the base 2 equal to 0.940 bits, so this is the meaning of our Info D . A log function to the base 2 is used because the information is encoded in bits.

So, Info (D) or entropy, another name for entropy is just the average amount of information needed to identify the class label of a tuple in D . Generally, the lesser the value of entropy that means we need very less informations to identify the class label of a tuple D , so generally the value of entropy should be less, so that attribute will be chosen for classification.

(Refer Slide Time: 22:44)

Attribute Selection Measures

- It is quite likely that the partitions will be impure (e.g., where a partition may contain a collection of tuples from different classes rather than from a single class).
- How much more information would we still need (after the partitioning) in order to arrive at an exact classification?
- This amount is measured by
$$\underline{Info}_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times Info(D_j)$$
- The term $|D_j| / |D|$ acts as the weight of the j^{th} partition. $Info_A(D)$ is the expected information required to classify a tuple from D based on the partitioning by A .

It is quite likely that the partitions will be impure, where a partition may contain a collection of tuples from different classes rather than from the single class, so how much more information would still need, after the partition in order to arrive an exact classification. So, this amount is measured by $Info(D)$ for one attribute that is for $\sum_{j=1}^v$ for all the splits; (modulus of (D_j) divided by modulus of (D)) multiplied by $Info(D_j)$, this is nothing but your entropy.

The term D_j ; modulus D_j divided by modulus D act as a weight of j^{th} operation, $Info(D)$ for attribute A is expected information required this one, expected information required to classify a tuple from D based on the partitioning by A .

(Refer Slide Time: 23:51)

Information Gain

- The smaller the expected information (still) required, the greater the purity of the partitions
- Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A). That is,

$$Gain(A) = Info(D) - Info_A(D)$$

- The attribute A with the **highest information gain**, ($Gain(A)$), is chosen as the splitting attribute at node N .

So, the information gain; the smaller the expected information required the greater the purity of the partitions, so information gain is defined as the difference between the original information requirement that is done by based on just proportion of classes and the new requirement that is obtained after partitioning on A. So, the Gain A = Info (D) - Info (D) for attribute A, I have used this example in my coming classes with the help of numerical example, I have explain how to find out the gain A.

The attribute A with the highest information gain is chosen as a splitting attribute at node N, you see that the entropy should be very smaller but the information gain should be higher for choosing an attribute.

(Refer Slide Time: 24:46)

Gini Index

- Gini index is used to measures the impurity of D, a data partition or set of training tuples, as

$$Gini(D) = 1 - \sum_{i=1}^m p_i^2$$

- Where p_i is the probability that a tuple in D belongs to class C_i and is estimated by $|C_{i,D}|/|D|$.
- The sum is computed over 'm' classes.
- The Gini index considers a binary split for each attribute

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

$$Gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$$

Next we will go to the next concept Gini index; Gini index is used to measure the impurity of D, the data partition or set of training tuples, the formula for Gini D = $1 - \sum_{i=1}^m p_i^2$, where p_i is the probability that tuples in D belongs to a class C_i and is estimated by modulus of ($C_i D$) divided by $|D|$, for example I will explain for this dataset, how to find out Gini index.

So, 1 minus; so how many yes is there here, it is 9 yes is there, so $(9 / 14)$ whole square minus, how many no is there; 5 no is there; $(5 / 14)$ whole square, so $1 - (9 / 14)^2 - (5 / 14)^2$ equal to 0.459, this is Gini index because the sum is computed over m classes, we are doing for

all the for m equal to 1, m equal to 2, the Gini index considers a binary split for each attribute. So, here we are going to get only binary split if you use Gini index.

(Refer Slide Time: 26:02)

Gini Index

- When considering a binary split, we compute a weighted sum of the impurity of each resulting partition
- For example, if a binary split on A partitions D into D_1 and D_2 , the Gini index of D given that partitioning is-

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

- For each attribute, each of the possible binary splits is considered
- For a discrete-valued attribute, the subset that gives the minimum Gini index for that attribute is selected as its splitting subset

When considering a binary split, we compute a weighted sum of impurity of each resulting partitions for example, if a binary split on a partitions D into 2 category; one is D1, D2, then the Gini index of D given that partitioning is; so Gini D for each attribute = (modulus of (D1) divided by |D1|). Gini of D1 + (modulus of (D2) divided by modulus of D). Gini of D2. So, in my coming lectures I have used this formula also with the help of a numerical example.

There you can have very clear understanding how we are finding this, for each attribute each of the possible binary split is considered, for a discrete valued attribute the subset that gives the minimum Gini index, you have to remember this, minimum Gini index for that attribute is selected as its splitting subset.

(Refer Slide Time: 27:02)

Gini Index

- For continuous-valued attributes, each possible split-point must be considered
- The strategy is similar where the midpoint between each pair of (sorted) adjacent values is taken as a possible split-point.
- For a possible split-point of A, D_1 is the set of tuples in D satisfying $A \leq$ split point, and D_2 is the set of tuples in D satisfying $A >$ split point.
- The reduction in impurity that would be incurred by a binary split on a discrete- or continuous-valued attribute A is

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

- The attribute that maximizes the reduction in impurity (or, equivalently, has the minimum Gini index) is selected as the splitting attribute

For continuous valued attributes, each possible split point must be considered; the strategy is similar where the midpoint between each of the pair, adjacent value is taken as a possible split point. If there is a continuous variable, the midpoint in a sorted dataset, the midpoint should be taken as the splitting criteria. For a possible split of A, D_1 is the set of tuples in D satisfying A less than or equal to split point.

And D_2 is the set of tuples in D satisfying A greater than split point, the reduction in impurity that would be incurred by a binary split on a discrete or continuous valued attribute A is delta of Gini A = Gini D, this is for our class variable - Gini D for a particular attributes. The attribute that maximises the reduction in impurity has the otherwise, which is having minimum Gini index is selected for splitting attribute. So, this value should be maximum otherwise, this will be maximum only if the Gini index is minimum.

(Refer Slide Time: 28:16)

Which attribute selection measure is the best?

- All measures have some bias.
- The time complexity of decision tree generally increases exponentially with tree height
- Hence, measures that tend to produce shallower trees (e.g., with multiway rather than binary splits, and that favour more balanced splits) may be preferred.
- However, some studies have found that shallow trees tend to have a large number of leaves and higher error rates
- Several comparative studies suggest no one attribute selection measure has been found to be significantly superior to others.

So, we have seen 2 methods; one method is information gain, another method is Gini index, there is one more method called gain ratio that I will take in my coming classes. So, how to choose which attribute method has to be chosen, all measures have some bias for example, this technique information gain also having some bias that I will explain in coming class. The time complexity of decision tree generally increases exponentially with the tree height.

Hence, measures that tend to produce shallower trees that is with multiway rather than binary split and that favour more balanced split may be preferred that is why most of the time Gini index is chosen because that is giving a balance split however, some studies have found that shallow trees tend to have a large number of leaves and higher error rates, several comparative studies suggest no one attribute selection measures has been found to be significantly superior to others.

(Refer Slide Time: 29:24)

Tree Pruning

- When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers
- Tree pruning use statistical measures to remove the least reliable branches
- Pruned trees tend to be smaller and less complex and, thus, easier to comprehend
- They are usually faster and better at correctly classifying independent test data than unpruned trees

Next, we will go the concept called tree pruning, when a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. So, tree pruning use statistical measures to remove the least branches, pruned trees tend to be smaller and less complex and thus easier to comprehend, they are usually faster and better at correctly classifying independent test data than unpruned trees.

(Refer Slide Time: 30:01)

How does Tree Pruning Work?

- There are two common approaches to tree pruning: **pre-pruning** and **post-pruning**.
- In the **pre-pruning** approach, a tree is “pruned” by halting its construction early (e.g., by deciding not to further split or partition the subset of training tuples at a given node).
- When constructing a tree, measures such as statistical significance, information gain, Gini index can be used to assess the goodness of a split.

How does tree pruning work; there are 2 common approaches to tree pruning; one is pre-pruning, another one is post-pruning. In the pre-pruning approach, the tree is pruned by halting its construction early that is by deciding not to further split or partition the subset of training

tuples at a given node, when constructing a tree measures such as statistical significance, information gain, Gini index can be used to assess the goodness of a split.

(Refer Slide Time: 30:34)

How does Tree Pruning Work?

- The **post-pruning** approach removes sub_trees from a “fully grown” tree
- A subtree at a given node is pruned by removing its branches and replacing it with a leaf
- The leaf is labelled with the most frequent class among the subtree being replaced
- For example, the subtree at node “A3?” in the unpruned tree of Figure 1.2
- The most common class within this subtree is “class B”
- In the pruned version of the tree, the subtree in question is pruned by replacing it with the leaf “class B”

Now, let us talk about the post-pruning; the post pruning approach remove the sub tree from a fully grown tree, a sub tree at a given node is pruned by removing its branches and replacing it with a leaf, so the leaf is labelled with the most frequent class among the sub tree being replaced.

(Refer Slide Time: 30:55)

How does Tree Pruning Work?

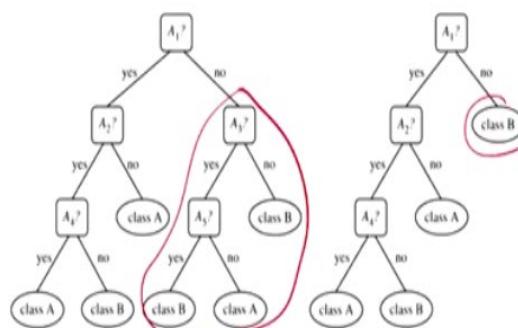


Figure: 1.2 An unpruned decision tree and a post-pruned decision tree

Look at this picture which is given in the next slide, assume that we are going to remove this portions, what is happening; in this when you look at this one, if you are removing this portion of the tree, the class B is frequently occurring, so we have to bring as a leaf node, in that leaf node,

the class B has to be retained. For example, the subtree at the node A3 in the unpruned tree is shown in figure 1.2.

The most common class within the subtree is class B, look at this there are class B is there, class B is there, class A is there, 1 class A is there, 2 class B is there, so the most common class with this subclass is class B. The pruned version of the tree, the subtree in the question in is pruned by replacing with the leaf class B, you see that this is the pruned version, in that we have retain class B. This figure explains unpruned decision tree and the post pruned decision tree.

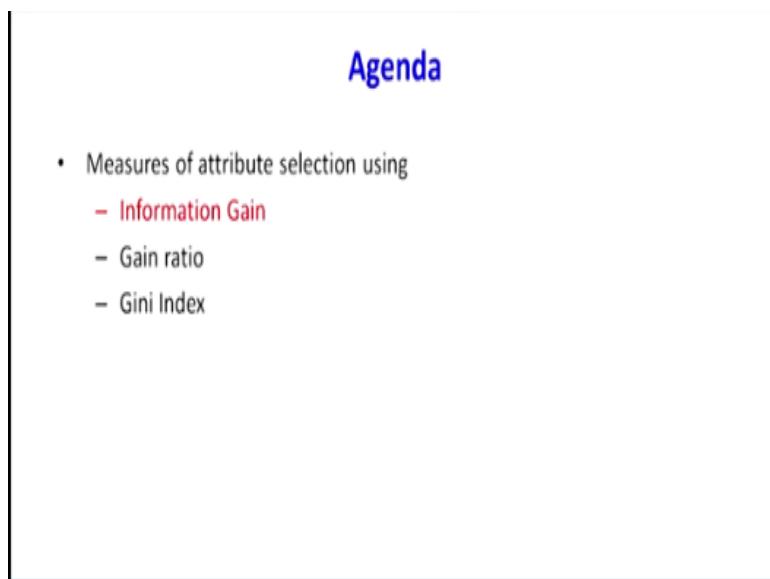
What you have done in this lecture; I have introduced what is classification regression tree CART model, then I have explained different terminology, which are frequently used in the CART model, then I have explained the theory behind different attribute selection measures like information gain, Gini index. At the end, I have explained how to do the pruning of the tree. In the next class with the help of a numerical example, I will explain how to do or how to select different attributes. For example, with the help of information how to choose attributes; correct attributes, thank you very much.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology - Roorkee

Lecture – 58
Measures of Attribute Selection

In our previous class, I have given introduction to classification regression tree, in this lecture I am going to take some example, some numerical examples; with the help of numerical examples, I am going to explain how to select attribute for the CART model.

(Refer Slide Time: 00:47)



Agenda

- Measures of attribute selection using
 - Information Gain
 - Gain ratio
 - Gini Index

The agenda for this lecture is measures of attribute selection using; there are 3 measures for selecting attributes. Here, what is the meaning of attribute is choosing an independent variables for making classification, so there are 3 criteria; one is we can choose attribute with the help of information gain, another measure is gain ratio, the third one is Gini index. In this lecture, I am going to take the first criteria that is information gain.

(Refer Slide Time: 01:24)

Example

- The following Table presents a training set, D, of class-labeled tuples randomly selected from the AllElectronics customer database

Han, J., Pei, J. and Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier.

<i>RID</i>	<i>age</i>	<i>income</i>	<i>student</i>	<i>credit.rating</i>	<i>Class: buys.computer</i>
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle.aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle.aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle.aged	medium	no	excellent	yes
13	middle.aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

With the help of this attribute, I am going to explain how to choose the attribute, now, taken one sample example, this example is taken from this book, Han and Kamber, the book title is Data mining, concepts and techniques. The problem is there are 1, 2, 3, 4, 5, there are 5 column is there, these 5 columns is called attributes, the last column is class that is a buys computer.

So, what this database says that the company; the database called all electronics customer database, they are going to see what kind of customers, they are going to buy the computer, they have 1 attributes or variable called age in that they have in different levels; youth, middle aged, senior. In income there are 3 levels; high, medium, low. In student; yes or no, in credit rating whether their credit rating is fair or excellent.

So, the final objective is we have to make a decision tree or classification tree for this dependent variable that is buy say computer, for choosing that one out of these 4 variables, we want to know from which variable we have to started. For that purpose, the information gain criteria is taken as a measure, let us see how it is working. The following table represents a training set, the data set called D of class labelled tuples randomly selected from all electronics customer database.

(Refer Slide Time: 03:13)

Example

- In this example, each attribute is discrete-valued
- Continuous-valued attributes have been generalized
- The class label attribute, buys computer, has two distinct values (namely, {yes, no}); therefore, there are two distinct classes (that is, $m = 2$)
- Let class C_1 correspond to 'yes' and class C_2 correspond to 'no'.
- There are nine tuples of class 'yes' and five tuples of class 'no'.
- A (root) node N is created for the tuples in D

The tuples is nothing but the full database called tuples, in this example each attribute is discrete valued, what are the attributes here; age is an attribute, it is a discrete because there is no continuous value, income is another attribute, student is another attribute, credit rating is another attribute and "buys computer" also an attribute, all are categorical variable there is no continuous variable here.

The class labels attributes that is in the last column the variable called buys computer has 2 distinct value namely yes, no therefore, there are 2 distinct classes, this m equal to 2, so this value m equal to 2 I am going to use in coming slides, please remember this m equal to; because there are 2 levels the person is going to buy the computer or not, let the class C_1 corresponds to yes, class C_2 corresponds to no.

There are 9 tuples of class yes and 5 tuples of class no, a root node N is created for the tuples in D, so for making this root node we have to find out which variable, from which variable you have to start this root node, 1, 2, 3, 4 out of this 4 variables; age, income, student and credit rating, we are going to find out which variable is going to be in the root node.

(Refer Slide Time: 04:36)

Expected information needed to classify a tuple in D

- To find the splitting criterion for these tuples, we must compute the information gain of each attribute
- Let us consider Class: buys_computer as decision criteria D
- Calculate information:
- $= -p_y \log_2(p_y) - p_n \log_2(p_n)$
- Where p_y is probability of 'yes' and p_n is probability of 'no'

$$Info(D) = -\frac{9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits.}$$

Here what we are going to do; expected information needed to classify a tuples D, to find the splitting criterion for these tuples, we must compute the information gain of each attributes. Let us consider the class, buys computer actually it is variable, buys underscore as a decision criteria D, so calculate informations that is $- p_y \log_2(p_y)$ to the base 2 $- p_n \log_2(p_n)$ to the base 2, where the p_y is a probability of yes and p_n is probability of no.

So, the another name for this equation is entropy, so Info D py, let us see what is a py; in the last column when you look at this how many yes is there; 1, 2, 3, 4, 5, 6, 7, 8, 9 yes is there, so 9 yes is there out of 14, $- (9/14) \log_2(9/14)$; again, $9/14$; (9 divided by 14) to the base 2 minus; then how many no's is there because out of 14, 9 is yes, so remaining 5 is no, (5 divided by 14) $\log_2(5/14)$.

(Refer Slide Time: 06:10)

Calculation of entropy for 'Youth'

- Age can be:

- youth
- Middle_aged
- Senior

- Youth

Youth	Class: buys computer
Yes	2
No	3

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

So, this is 0.940 bits for our dependent variable, now let us calculate entropy for the variable age, so age can be; see there are 3 levels in age, one is youth, second one is middle-aged, third one is senior. If you take youth how many people has given yes is their option, so here youth there is 1, youth because here also yes, yes there are 2 people, so youth how many people are yes; 2. So, how many people say no; this one, 1, 2, 3, so out of 5 youth, 2 youth they told their yes; yes means yes for buying computer, 3 youth told no for buying computer.

(Refer Slide Time: 07:08)

Calculation of entropy for 'Youth'

- Calculate Entropy for youth:

$$\text{Entropy}_{\text{youth}} = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5}$$

- Middle_aged

middle	Class: buys computer
Yes	4
No	0

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle age	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle age	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle age	medium	no	excellent	yes
13	middle age	high	yes	fair	yes
14	senior	medium	no	excellent	no

Let us calculate entropy for youth, so out of 5 entropy for youth is, this entropy for youth equal to $-2 / 5 \log_2 2 / 5$ divided by 5 to the base 2 - 3 people have told no, so $-3 / 5 \log_3 3 / 5$ divided by 5 to the base 2, so this is entropy for youth. Then we will go to the next level, the next level is middle aged; in this middle aged, how many people told yes? 1, this 2, middle aged yes, middle aged yes, so there are 4, 4 people told yes, so no is 0.

(Refer Slide Time: 08:07)

Calculation of entropy for 'Middle Age'

- Calculate Entropy for middle_aged

$$= -\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4}$$

$$= 0$$

- For Senior

Senior	Class: buys computer
Yes	3
No	2

ID	age	income	student	credit_rating	Class: buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Because only there are 4; 1, 2, 3, 4 middle aged people, so calculate entropy for a middle aged here, entropy - 4 divided by 4 log (4 by 4) to the base 2 - 0 by 4 log 2 0 divided by 4, so here entropy is 0. Similarly, the next level is senior, when it is senior how many people told yes, so this senior is yes, this senior is yes, this senior is yes, so there are 3 people told yes, so how many senior level they told no; this they told no here and this senior also told no, there are 2 people, so out of 5, 3 people told yes for buying computer, 2 people have told no for buying computer.

(Refer Slide Time: 08:59)

Calculate Entropy for senior

$$\begin{aligned} &\text{Calculate Entropy for senior} \\ &= -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \end{aligned}$$

The expected information needed to classify a tuple in D if the tuples are

partitioned according to age is

$$\begin{aligned} \text{Info}_{age}(D) &= \frac{5}{14} \times \left(-\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \right) \\ &\quad + \frac{4}{14} \times \left(-\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4} \right) \\ &\quad + \frac{5}{14} \times \left(-\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) \\ &= 0.694 \text{ bits.} \end{aligned}$$

Let us find out entropy for senior, so - 3 by 5 log 2 log 3 divided by 5 to the base 2 - 2 by 5, remaining 2 people told no - 2 divided by 5 log 2 divided by 5 to the base 2. So, now what you are going to see, the expected information needed to classify a tuple in D, if the tuples are

partitioned according to age is; so what we are going to do, we are going to find out the expected information needed.

There are 5 element right, there are 5, how we got this 5; for example, youth is 5, so it is 5 by 14, then middle aged 4 divided by 14, then senior 5 out of 14, so now we are finding the expected information needed that is 0.694 bits.

(Refer Slide Time: 10:05)

Calculate Entropy for senior

$$\begin{aligned} &\text{Calculate Entropy for senior} \\ &= -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \end{aligned}$$

Calculation of entropy for senior category; so we have seen in our previous slide out of 5, there are 5 senior score is there, out of 5, 3 people have answered yes, 2 people have answered no, so the entropy is – 3 divided by 5 log 3 divided by 5 to the base 2 - 2 divided by 5 log 2 by 5 to the base 2. So, now we have got the entropy for all the levels.

(Refer Slide Time: 10:36)

The expected information needed to classify a tuple in D according to age

The expected information needed to classify a tuple in D if the tuples are

$$\begin{aligned} \text{partitioned according to age is } &Info_{age}(D) = \frac{5}{14} \times \left(-\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \right) \\ &+ \frac{4}{14} \times \left(-\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4} \right) \\ &+ \frac{5}{14} \times \left(-\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) \\ &= 0.694 \text{ bits.} \end{aligned}$$

Now, let us find the expected information needed to classify a tuple in D, according to age so, the expected information needed to classify a tuple in D, if the tuples are partitioned according to age is, so here nothing but we are finding weighted, so because there was a 5 youth out of 14, there are 4 middle-aged people out of 14, there are 5 senior out of 14, so 5 divided by 14, then corresponding entropy, 4 divided by 14 corresponding entropy, 5 divided by 14 corresponding entropy, so it is 0.694 bits,.

(Refer Slide Time: 11:24)

Calculation information Gain of Age

- Gain of Age:

$$Gain(age) = \underline{Info(D)} - \underline{Info_{age}(D)} = 0.940 - \underline{0.694} = 0.246 \text{ bits.}$$

Now, calculation of information gain of age, so the gain of age is; see that gain of age = Info D – Info age D for only 1 variable age, so this Info D 0.940 which we have found for the class attributes that is the our dependent variable, this is only for the info variable age. So, the difference is 0.246 bits.

(Refer Slide Time: 11:58)

Calculation information Gain of Income

- Calculation of gain for income:

- Income can be:
 - High
 - Medium
 - Low

RID	age ✓	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle.age	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle.age	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle.age	medium	no	excellent	yes
13	middle.age	high	yes	fair	yes
14	senior	medium	no	excellent	no

Now, we will go to next variable, we have seen for age, now we will go for variable income. In the income, the same way we will repeat the procedure, in the income there are 3 levels there; high, medium, low, so we will find the entropy for high, medium and low, then we will find the expected information required, then we will find the information gain, how will you find the information gain? So, it is the gain from our dependent variable minus this income variable, let us find out that one.

(Refer Slide Time: 12:31)

Calculate Entropy for high

- High :

High	Class: buys computer
Yes	2
No	2

- Calculate Entropy for high:

$$= -(2/4)\log_2(2/4) - (2/4)\log_2(2/4)$$

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle.age	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle.age	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle.age	medium	no	excellent	yes
13	middle.age	high	yes	fair	yes
14	senior	medium	no	excellent	no

So, when you go for high income, how many people have told yes? Here that is 1, here yes, then how many people told no, when the level is this 1, 2 that is all, so out of there are 4 values under the income level, out of 4, 2 have answered yes for buying computer, 2 have responded no for buying computer. So, we will find out entropy for high, so $-2/4 \log_2(2/4)$ divided by 4 log 2 divided by 4 to the base 2 – 2 divided by 4 log 2 divided by 4 to the base 2.

(Refer Slide Time: 13:20)

Calculate Entropy for 'medium'

- Medium:

Medium	Class: buys computer	RID	age	income	student	credit_rating	Class: buys.computer
Yes	4	1	youth	high	no	fair	no
		2	youth	high	no	excellent	no
No	2	3	middle.age	high	no	fair	yes
		4	senior	medium	no	fair	yes
		5	senior	low	yes	fair	yes
		6	senior	low	yes	excellent	no
		7	middle.age	low	yes	excellent	yes
		8	youth	medium	no	fair	no
		9	youth	low	yes	fair	yes
		10	senior	medium	yes	fair	yes
		11	youth	medium	yes	excellent	yes
		12	middle.age	medium	no	excellent	yes
		13	middle.age	high	yes	fair	yes
		14	senior	medium	no	excellent	no

So, this value we got now, we will go to the next category level. The next category level is medium; in the medium, we are going to find out how many people are answered yes, medium yes, medium no, this is no, then medium yes, medium yes, medium yes, 1, 2, 3, 4, so there are 4 people have answered yes for buying computer. So, let us see in medium how many people answered no.

So, medium no, then medium this one, medium no, so 2 people have answered no for buying computed, so we will find the entropy, so – 4 divided by 6 log 4 divided by 6 to the base 2 – 2 divided by 6 log 2 divided by 6 to the base 2, so we will get an entropy for medium.

(Refer Slide Time: 14:11)

Calculate Entropy for 'low'

- Low :

Low	Class: buys computer	RID	age	income	student	credit_rating	Class: buys.computer
No	1	1	youth	high	no	fair	no
		2	youth	high	no	excellent	no
Yes	3	3	middle.age	high	no	fair	yes
		4	senior	medium	no	fair	yes
		5	senior	low	yes	fair	yes
		6	senior	low	yes	excellent	no
		7	middle.age	low	yes	excellent	yes
		8	youth	medium	no	fair	no
		9	youth	low	yes	fair	yes
		10	senior	medium	yes	fair	yes
		11	youth	medium	yes	excellent	yes
		12	middle.age	medium	no	excellent	yes
		13	middle.age	high	yes	fair	yes
		14	senior	medium	no	excellent	no

Then, we will find out the entropy for low; in low, how many yes is there; low yes, then low yes, then here low yes, so 3 people have answered yes for buying computer, how many

people are answered no, when they are low yeah, here it is there, this low no, so only 1 people answered no. So, out of 4, 3 answered yes for buying computer, 1 answered no for buying computer. Now, we will find out the entropy; - 1 divided by 4 log 1 divided by 4 to the base 2 - 3 divided by 4 log 3 divided by 4 to the base 2.

(Refer Slide Time: 15:03)

Gain of income

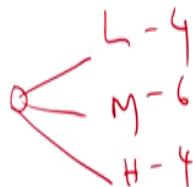
- The expected information needed to classify a tuple in D if the tuples are partitioned according to income is:
- $$\text{Info}_{\text{income}}(D) = \frac{4}{14} \left(-\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} \right) +$$

$$\frac{6}{14} \left(-\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} \right) +$$

$$\frac{4}{14} \left(-\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} \right)$$

$$= 0.911$$

$$\begin{aligned}\text{Gain of income : } &\text{Info}(D) - \text{Info}_{\text{income}}(D) \\ &= 0.94 - 0.911 = 0.029\end{aligned}$$



The expected information needed to classify a tuple in D if the tuples are partitioned according to income is; so we are finding this weightage, it is nothing but the weighted mean of the entropy, so 4 divided by 14, then it is 6 divided by 14, what is this 6; we will go back, medium will be 1, 2, 3, 4, 5, 6, so there are 6 medium. So, what is happening, it is like this; low, medium, high.

In low, how many variable is there; 1, 2, 3, 4; 4, in medium 1, 2, 3, 4, 5, 6; 6, in high 1, 2, 3, 4, so there are total 14, so 4 divided by 14 the entropy for low, then 6 divided 14, this 6 divided by 14, then the entropy plus 4 divided by 14 that entropy, the weighted entropy is 0.911. So, the information gain of income variable is; so Info D that is for our dependent variable, Info income D is this 0.91, so the difference is 0.029. So, this represents expected information needed to classify a tuple D, if the tuples are partitioned according to income.

(Refer Slide Time: 16:35)

Calculation of gain for student

- Calculation of gain for student

- Student can be:

- Yes (7)
- No (7)

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle.aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle.aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle.aged	medium	no	excellent	yes
13	middle.aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

We will go to the next variable is a student, in student there are 2 level is there; one is yes and no. So, how many yes is there? 1, 2, 3, 4, 5, 6, 7; 7 yes is there. How many no is there? 1, 2, 3, 4, 5, 6, 7, so 7 no is there. Now, we will find out the entropy when it is yes, we will find out the entropy when it is no.

(Refer Slide Time: 17:08)

Calculate Entropy for No

- No :

No	Class: buys computer
Yes	3
No	4

- Calculate Entropy for No:

$$= -(3/7)\log_2(3/7) - (4/7)\log_2(4/7)$$

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle.aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle.aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle.aged	medium	no	excellent	yes
13	middle.aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

When it is no, how many yes, so here one is there, no, yes, no, no, no, yes, so out of 7, 3 people have answered yes to buy the computer. So, how many people answered no for buying computer when they it is no, so this is 1, 2, 3, here one more is there 4, there is a 4. So, entropy for no is - 3 divided by 7 log 3 divided by 7 to the base 2 - 4 divided by 7 log 4 divided by 7 to the base 2.

(Refer Slide Time: 18:03)

Calculate Entropy for 'Yes'

- Yes :

Yes	Class: buys computer
Yes	6
No	1

- Calculate Entropy for Yes:

$$= -(6/7)\log_2(6/7) - (1/7)\log_2(1/7)$$

RID	age	income	student	credit.rating	Class: buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle.aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle.aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle.aged	medium	no	excellent	yes
13	middle.aged	high	no	fair	yes
14	senior	medium	no	excellent	no

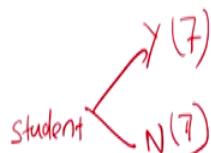
Now, we will find out entropy for yes, so when it is yes, how many people are answered yes for buys a computer; 1, 2, 3, 4, 5, 6, so there are 6 yes is there. Now, how many no; when there is yes, how many people have answered no, here this one, there is only 1 no is there, so the entropy for yes is - 6 divided by 7 log 6 divided by 7 to the base 2 - 1 divided by 7 log 1 divided by 7 to the base 2.

(Refer Slide Time: 18:44)

Gain of student

- The expected information needed to classify a tuple in D if the tuples are partitioned according to student is:

$$\begin{aligned} \text{Info}_{\text{Student}}(D) &= (7/14)(-(3/7)\log_2(3/7) - (4/7)\log_2(4/7)) + \\ &\quad (7/14)(-(6/7)\log_2(6/7) - (1/7)\log_2(1/7)) \\ &= 0.789 \end{aligned}$$



- Gain(student) :

$$\begin{aligned} \text{Info}(D) - \text{Info}_{\text{student}}(D) \\ = 0.94 - 0.789 = 0.151 \end{aligned}$$

Now, we will find out expected information, so for the expected informations, so what we have to do; we have to find out the weighted entropy. So, the weighted entropy is 7 divided by 14 because already we have seen when there was a student, how many yes is there, how many no is there? There are 7 yes, there are 7 no, so out of 14, 7 divided by 14 and that is corresponding entropy plus for no, there is 7 divided by 14 corresponding entropy.

So, this was the expected information needed, so the gain is; so 0.94 for our dependent variable and for this student variable, it is 0.789, so we got this one the gain; the gain is 0.151.

(Refer Slide Time: 19:41)

Calculation of gain for credit rating

- Calculation of gain for credit rating
- Credit rating can be:
 - Fair - 8
 - Excellent - 6

RID	age	income	student	credit rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Next, we will go to the next variable; credit rating. In credit rating, how many levels is there? Fair, excellent, there are 2 level is there, fair and excellent. So, how many fair is there? 1, 2, 3, 4, 5, 6, 7, 8. How many excellent is there; 1, 2, 3, 4, 5, 6, out of 14, 6 is there. So, now we will find out the entropy, when it is a fair, we will find out entropy when it is excellent for the credit rating.

(Refer Slide Time: 20:14)

Calculate Entropy for Fair

- Fair :

Fair	Class: buys computer
Yes	6
No	2

- Calculate Entropy for Fair:

$$= -(6/8)\log_2(6/8) - (2/8)\log_2(2/8)$$

RID	age	income	student	credit rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

So, for fair how many people are answered yes; fair, fair, fair yes, fair yes, fair yes, fair yes, fair yes, fair yes, 1, 2, 3, 4, 5, 6. So, how many people are fair and same time it is no, so this

fair no, this fair no, there are 2. So, how to find out the entropy for fair; - 6 divided by 8 log of 6 divided by 8 to the base 2 - 2 divided by 8 log 2 divided by 8 to the base 2.

(Refer Slide Time: 21:00)

Calculate Entropy for Excellent

- Excellent :

Yes	Class: buys computer	RID	age	income	student	credit.rating	Class: buys.computer
Yes	3	1	youth	high	no	fair	no
Yes	3	2	youth	high	no	excellent	no
No	3	3	middle aged	high	no	fair	yes
		4	senior	medium	no	fair	yes
		5	senior	low	yes	fair	yes
		6	senior	low	yes	excellent	no
		7	middle aged	low	yes	excellent	yes
		8	youth	medium	no	fair	no
		9	youth	low	yes	fair	yes
		10	senior	medium	yes	fair	yes
		11	youth	medium	yes	excellent	no
		12	middle aged	medium	no	excellent	yes
		13	middle aged	high	yes	fair	yes
		14	senior	medium	no	excellent	no

Now, let us find the entropy for excellent, so how many people are yes for buying compute, when they are excellent; excellent 1, excellent 2, excellent 3, there are 3 people. How many people told no; excellent no, excellent no, excellent no, so 3 people are answered yes for buying computer when their level is excellent, 3 people are answered no for buying computer when their level is excellent. So, the entropy for excellent is - 3 divided by 6 log 3 divided by 6 to the base 2 - 3 divided by 6 log 3 divided by 6 to the base 2.

(Refer Slide Time: 21:44)

Gain for credit rating

- The expected information needed to classify a tuple in D if the tuples are partitioned according to Credit rating is:
- $\text{Info}_{\text{Credit rating}}(D) = (8/14) \left(-(6/8)\log_2(6/8) - (2/8)\log_2(2/8) \right) + (6/14) \left(-(3/6)\log_2(3/6) - (3/6)\log_2(3/6) \right)$
 $= 0.892$

- Gain for credit rating :

$$\text{Info}(D) - \text{Info}_{\text{Credit rating}}(D)$$

$$= 0.94 - 0.892 = 0.048$$

Now, we will find out the expected information needed to classify in a tuple D, if the tuples are partitioned according to credit rating. So, here in the credit rating there are 2 levels; one is

as I told you one is fair, another one is excellent. So, for fair there are 8 items, so 8 divided by 14 and corresponding their entropy, the remaining 6 divided by 14 and their corresponding their entropy.

So, the expected information needed is 0.892, now we will find out gain for credit rating. What is the meaning of gain for credit rating? If we use credit rating as the root variable, how much information is required, so this variable we got it when it is 0.94 for the dependent variable, this is for credit rating variable just now we got it 0.892, so the difference is 0.048.

(Refer Slide Time: 22:52)

Independent variable	Information gain
Age	0.246
Income	0.029
Student	0.151
Credit_rating	0.048

Now, I have summarized; if we use age as the classifier, the information gain is 0.246, if we use income as the classifier, the information gain is 0.029, if we use student as a classifier the information gain is 0.151, if we use credit rating as a classifier the information gain is 0.048. So, out of this 4, the highest value is 0.246, so now we will start keeping age as a classifier variable, so that is application of this.

(Refer Slide Time: 23:36)

Selection of root classifier

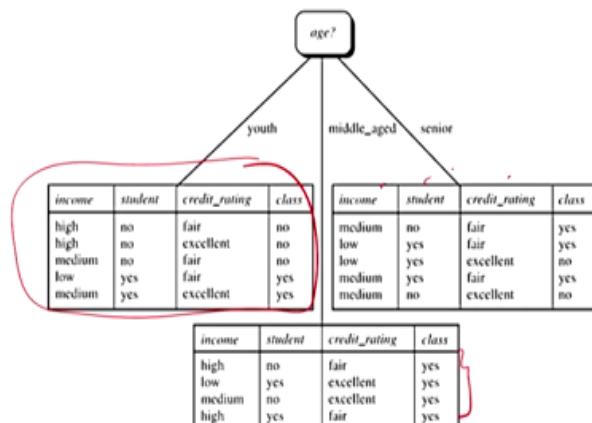
- Because age has the highest information gain among the attributes, it is selected as the splitting attribute
- Node N is labelled with age, and branches are grown for each of the attribute's values
- The tuples are then partitioned accordingly
- Notice that the tuples falling into the partition for age = middle aged all belong to the same class
- Because they all belong to class "yes," a leaf should therefore be created at the end of this branch and labelled with "yes."

So, what happened because age has the highest information gain among the attributes, it is selected as the splitting attribute, node N is labelled with age and the branches are grown for each of the attributes values. The tuples are then partitioned accordingly, notice that the tuples falling into the partition for age, when there is a middle aged all belongs to the same class.

So, we need not go for further classification because all are belongs to; all middle aged people are answered yes for buying computer, so further classification is not required because they are belongs to class yes, a leave should therefore be created at the end of this branch and labelled with yes.

(Refer Slide Time: 24:23)

Decision tree



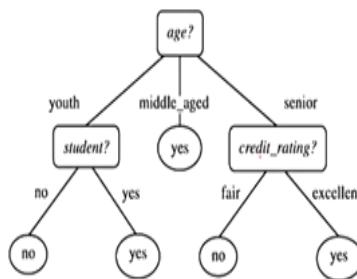
This was the decision tree, so what happened; the age is the classifier, if there is age; youth, middle aged and senior, there are 3 level is there. So, for middle aged all are answering yes, so further classification is not required. So, if it is youth there are some other variables is there, income is there, student is there, credit rating is there, so the information gain algorithm methodology which we have used can be used this group also.

Again, we can find out, out of these 3 variables; income, student, credit rating, which variable should appear here, same way the one classifier is a senior, when it is a senior there are 1, 2, 3 variable is there, you can find out this information criteria, so out of these 3 variable which variable should come into the this node, so this way we can continue our classification procedure.

(Refer Slide Time: 25:25)

Decision tree

- The final decision tree returned by the algorithm is shown in Figure



This was our final decision tree returned by algorithm I shown in the figure, we started with the age; when you started with the age there are 3 level was there in the age; youth, middle aged, senior, so we are stopping because all are yes there, we do further classification required. Then you see that as I told you previously there are 3 options income, student, credit rating, so out of this, the student yes appeared one classifier on the left hand side.

So, when there is a student we can there is a 2 possibility; yes or no and the right hand side when it is seen here, you see that there are 3 possibilities there, we can classify with respect to income, student, credit rating. So, what happened is student already we have done that one, so now we go for credit rating, this also got by using that information gain measures, so this was our final decision tree.

We have seen different measures for selecting the attributes, there are 3 measure is there; one measure is information gain, another one is gain ratio, another one is gain index. In this lecture, what I have done using information gain as a measure by taking one numerical example, I have explained how to choose an attribute. In the next class, I will take another measures for choosing the attribute that is gain ratio and Gini index with that example, we will continue in my next lecture, thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology – Roorkee

Lecture – 59
Attribute Selection Measures in CART – II

In this lecture, we are going to see how to select attributes in CART model. In our previous lectures, we have seen how to choose attributes by using gain value. In this lecture, there are another two methods for choosing the attributes. One is gain ratio; another one is Gini index. This lecture we are going to see other two criteria for choosing the attributes.

(Refer Slide Time: 00:51)

Gain Ratio

- The information gain measure is biased toward tests with many outcomes
- That is, it prefers to select attributes having a large number of values
- For example, consider an attribute that acts as a unique identifier, such as product ID.
- A split on product ID would result in a large number of partitions (as many as there are values), each one containing just one tuple

First, we will look at the Gain ratio. The information gain measure is biased towards tests with many outcomes. That is, it prefers to select attributes having a large number of values. For example, consider an attribute that act as a unique identifier, such as product ID.

(Refer Slide Time: 01:41)

Gain Ratio

- Because each partition is pure, the information required to classify dataset D based on this partitioning would be $\text{Info}_{\text{product ID}}(D) = 0$
- Information Gain = $\text{Info } D - \text{Info}_{\text{product ID}}(D)$ = maximum
- Therefore, the information gained by partitioning on this attribute is maximal
- Clearly, such a partitioning is useless for classification
- Gain ratio is an extension to information gain which attempts to overcome this bias

A split on product ID would result in a large number of partitions as many as there values each one containing just one tuple because each partition is pure, the information required to classify dataset D based on this partitioning would be Info product ID will be 0, because there might be only one value in each partition. So Information gain is we know the formula, the formula for information gain is Info D – Info product ID for D because this value is going to be 0, so the information gain will be maximum.

Therefore, the information gained by partitioning on this attribute is maximal. Clearly, such partitioning is useless for classification, because there are going to be one element for each partition. So, the gain ratio is an extension to information gain, which attempts to overcome this bias.

(Refer Slide Time: 02:14)

Split information

- It applies a kind of normalization to information gain using a “split information” value defined analogously with $\text{Info}(D)$ as:

$$\text{SplitInfo}_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|} \right)$$

- D_j = single partition
- D = Data set
- This value represents the potential information generated by splitting the training data set, D, into v partitions, corresponding to the v outcomes of a test on attribute A

Let us see what is split information. It applies a kind of normalization to information gain using a split information value defined analogously with Info D. So, how to find out the split info for level for $D = - \sum_{j=1}^v v$ is different levels; ($|D_j|$ divided by $|D|$), multiplied by $\log_2 (|D_j| / |D|)$. Here the D_j is the single partition. In the next example, I will explain the D_j . D is dataset. So, this value that is the split info represents the potential information generated by splitting the training data set D into v partitions, corresponding to the v outcomes of test on attribute A.

(Refer Slide Time: 03:11)

Gain ratio

- Gain ratio differs from information gain, which measures the information with respect to classification that is acquired based on the same partitioning
- The gain ratio is defined as

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

- The attribute with the maximum gain ratio is selected as the splitting attribute

So, let us see the formula for the gain ratio. The gain ratio differs from the information gain, which measures the information with respect to classification that is acquired based on the same partitioning. So, the gain ratio is $(Gain(A)) / (SplitInfo(A))$. The attribute with the maximum gain ratio is selected as the splitting attribute. I have an example; with that I can explain how to use this gain ratio for choosing the attribute.

(Refer Slide Time: 03:43)

Gain Ratio example

- Consider the previous example for computation of gain ratio for the attribute income
- A test on income splits the data of the following Table into three partitions, namely low, medium, and high, containing four, six, and four tuples, respectively

Han, J., Pei, J. and Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier.

RID	age	income	student	credit_rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle.aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle.aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle.aged	medium	no	excellent	yes
13	middle.aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Look at this example. In my previous lecture, I was showing this dataset. Consider in our previous example for computation of gain ratio for the attribute income. So, we are going to take this variable first. We are going to find out gain ratio. I am going to say the procedure for only one attribute, like that you have to try for age attribute, for student attribute, for credit rating attribute.

Whichever is high that variable should be chosen for classification. A test on income splits the data into following table into three partitions. See, when you look at this income column, there are three levels, one is low, medium, and high containing in low, there are 6 element is there, in medium 6 is there, in high there is 4. This example is taken from the book data mining concepts and techniques.

(Refer Slide Time: 05:04)

Calculate Entropy for 'low'

- Low :

Low	Class: buys computer
Yes	3
No	1

- Calculate Entropy for Low:

$$= - (3/4)\log_2(3/4) - (1/4)\log_2(1/4)$$

RID	age	income	student	credit_rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle.aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle.aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle.aged	medium	no	excellent	yes
13	middle.aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Now, let us calculate the Entropy for level low. In level low, how many have answered yes to by computer when it is low. So how many people have answered low. When the level is low, so the entropy is $-(3 / 4) \log 3 / 4$ to the base 2 $-(1 / 4) \log 1 / 4$ to the base 2.

(Refer Slide Time: 05:44)

Calculate Entropy for buying class D

- Calculate information:
- $= -p_y \log_2(p_y) - p_n \log_2(p_n)$
- Where p_y is probability of yes and p_n is probability of no

$$Info(D) = -\frac{9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits.}$$

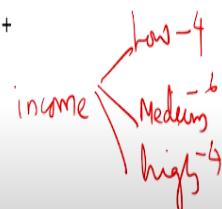
For finding the information gain for the attribute income, first we need to know the entropy for the class D. We know that how to find the entropy for the class D. It is $- p_y \log p_y$ to the base 2 $- p_n \log p_n$ to the base 2. So, the p_y is in our class, how many people have answered yes. When you look at the table, there was 9 yes and there is 5 no. So, for s it is $9 / 14$, $\log 9 / 14$ to the base 2 minus no. There is 5 no, so it is 5 no, so $5 / 14$, $\log 5 / 14$ to the base 2 equal to 0.940 bits.

(Refer Slide Time: 06:43)

Gain of income

- The expected information needed to classify a tuple in D if the tuples are partitioned according to income is:
- $Info_{income}(D) = (4/14)(-(2/4)\log_2(2/4) - (2/4)\log_2(2/4)) +$
 $(6/14)(-(4/6)\log_2(4/6) - (2/6)\log_2(2/6)) +$
 $(4/14)(-(1/4)\log_2(1/4) - (3/4)\log_2(3/4))$
 $= 0.911 \text{ bits}$

$$\begin{aligned} \text{Gain of income : } & Info(D) - Info_{income}(D) \\ & = 0.94 - 0.911 = 0.029 \end{aligned}$$



Now, let us find out the information gain for the attribute income. What is the meaning is, if you use income as an attribute for the classification, how much information you can gain. So the expected information needed to classify a tuple in D if the tuples are partitioned according to income is, so Info for attribute income is $4 / 14$, we got this 4 for the attribute income, there were three levels low, medium, and high.

In low, there was a 4, in medium there was 6, and in high there was 4 values. So, it is a kind of weighted attributed, because it is nothing but expected information needed. So, it is nothing but our weighted entropy. So, the weighted entropy is 0.911 bits. So, the gain of income is $\text{Info } D - \text{Info for the attribute income}$. So, we got this 0.94, which we got from this value, $0.94 - 0.911 = 0.029$.

(Refer Slide Time: 8:08)

Gain-Ratio(income)

- Calculation of split ratio:

$$\begin{aligned} \text{SplitInfo}_A(D) &= -\frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) - \frac{4}{14} \times \log_2\left(\frac{4}{14}\right) \\ &= 0.926. \end{aligned}$$

- Therefore, $\text{Gain-Ratio(income)} = 0.029 / 0.926 = 0.031$

Now, we will go for split ratio. So, the split ratio is equal to $-(4 / 14) \times \log_2(4 / 14)$ base 2 $- (6 / 14) \times \log_2(6 / 14)$ to the base 2 $- (4 / 14) \times \log_2(4 / 14)$ to the base 2. So, split info is 0.926. Therefore, the Gain-Ratio for the attribute income is $0.029 / 0.926$ is equal to 0.031.

(Refer Slide Time: 09:05)

Interpretation

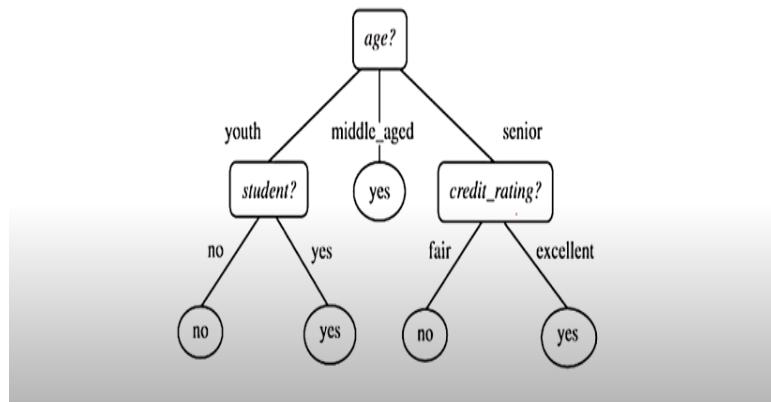
- Further we calculate the same for the rest 3 criteria (age, student, credit rating)
- The one with maximum Gain ratio value will result in the maximum reduction in impurity of the tuples in D and is returned as the splitting criterion



Further, we calculate the same for the rest of 3 criteria, what are the other three attributes, age is there, student is there, credit rating is there. The one with the maximum Gain ratio value will result in maximum reduction in impurity of the tuples in D and is returned as the splitting criterion. So, what we have to do, we have seen for one attribute that is income. There are another 3 attributes, such as age, student and credit rating. For these attributes also, we have to find out the information gained ratio.

That corresponding attribute should be taken as the splitting criterion. We have found the Gain ratio for the attribute income. The same way, there are other attributes like age, student and credit rating. For these attributes also, we have to find out the Gain ratio. One with the maximum gain ratio value will result in maximum reduction and impurities of the tuples in D and is written as the splitting criterion.

(Refer Slide Time: 10:09)



For example, assume that the attribute age is having maximum gain ratio, so that variable should be chosen as the splitting variable. Then from there, if there is a student, assume that the attribute age is having highest gain ratio, so that age is taken as the splitting variable. Then, there is a student. There are remaining other variables, for example, student, credit rating and so on. So, out of these, again you have to find the Gain ratio. So, out of this, which one is giving the maximum Gain ratio, that should be taken as the splitting criterion.

(Refer Slide Time: 10:50)

Decision tree using Gini index

- Let's take the introduction of a decision tree using Gini index
- Let D be the training data of the following table

Han, J., Pei, J. and Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier.

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Next, we will go to another criterion for choosing the attribute, that is Gini index. Let us take the introduction of decision tree using Gini index. Let D be the training data of the following table. So, this data also taken from the book data mining, concepts and techniques, the source is given here. Now, we are going to see how to find out the Gini index.

(Refer Slide Time: 11:12)

Example

- In this example, each attribute is discrete-valued
- Continuous-valued attributes have been generalized
- The class label attribute, buys computer, has two distinct values (namely, {yes, no}); therefore, there are two distinct classes (that is, $m = 2$)
- Let class C_1 correspond to 'yes' and class C_2 correspond to 'no'.
- There are nine tuples of class 'yes' and five tuples of class 'no'.
- A (root) node N is created for the tuples in D

In this example, each attribute is discrete-valued because all are in different category, so continuous-valued attributes have been generalized. We did not take continuous value. The class label attribute, buy computer, has two distinct values, yes or no, therefore, there are two distinct classes, $m = 2$. Let class C_1 correspond to yes and class C_2 correspond to no. There are nine tuples of class yes and five tuples of class no. A root n is created for the tuples D.

(Refer Slide Time: 12:07)

Calculation of Gini(D)

- We first use the following Equation for Gini index to compute the impurity of D:

$$\begin{aligned} \text{Gini}(D) &= 1 - \sum_{i=1}^m p_i^2, \\ &= \text{Gini}(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459. \end{aligned}$$

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle.aged	high	no	fair	yes✓
4	senior	medium	no	fair	yes✓
5	senior	low	yes	fair	yes✓
6	senior	low	yes	excellent	no
7	middle.aged	low	yes	excellent	yes✓
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes✓
10	senior	medium	yes	fair	yes✓
11	youth	medium	yes	excellent	yes✓
12	middle.aged	medium	no	excellent	yes✓
13	middle.aged	high	yes	fair	yes✓
14	senior	medium	no	excellent	no

Now, we go for calculation of Gini index. We first use the following equation for Gini index to compute the impurity of D. So, first we will find out the Gini of class D is $1 - \sum_{i=1}^m p_i^2$, (m is number of levels), p_i^2 . So, what is the p , how many s is there. So, $1 - (9 / 14)^2$, how many no is there, next level, so the Gini of class D = $1 - (9/14)^2 - (5 / 14)^2$, that is 0.459.

(Refer Slide Time: 12:54)

Gini index for income attribute

- Lets calculate Gini index for income attribute
- To find the splitting criterion for the tuples in D, we need to compute the Gini index for each attribute
- Let's start with the attribute income and consider each of the possible splitting subsets
- Income has three possible values, namely {low, medium, high}, then the possible subsets are {low, medium, high}, {low, medium}, {low, high}, {medium, high}, {low}, {medium}, {high}, and {}
 - Power set and empty set will not be used for splitting

Let us calculate, Gini index, previously found the Gini, the Gini index for income attribute. To find the splitting criterion for the tuples in D, we need to compute the Gini index for each attribute. In this, you take an example, income. Let us start with the attributes income and consider each of the possible splitting subsets. Incomes has three possible values, namely low, medium, high.

The possible subsets are low, medium, high, then all possible combinations low and medium, low and high, medium and high, then values which is having one value in the set, low, medium, high and null set. Power set and empty set will not be used for splitting. What is the power set where all the element is there, for example, low, medium, high is the power set, the null set is nothing but the empty set.

(Refer Slide Time: 14:00)

Gini index for income attribute

- Consider the subset{low, medium}
- This would result in 10 tuples in partition D1 satisfying the condition "income \in {low, medium}"
- The remaining four tuples of D (high) would be assigned to partition D2

(W) D1 (H) D2

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

So this will not be used for splitting. Now, we are going to split into two category. One is subset low and medium, because it is a binary classification, when you go for low and medium, another group. Suppose, there are two group is there, group 1 and group 2, in group 1, we have taken low and medium, obviously another set will be high. So, this would result in 10 tuples in partition D1.

We have to count how many low medium is there. We have got 10 tuples in partition D1, so this group is D1, that is why the condition income in the set low and medium. The remaining four tuples of D, the remaining is high, that is the remaining 4 that is D2, the remaining four tuples of D would be assigned to partition D2. What has happened, we have made a two subset, one is low medium and the remaining one is high.

(Refer Slide Time: 15:09)

Tuples in partition D1

- Low + Medium:

Medium + Low	Class: buys computer	RID	age	income	student	credit.rating	Class: buys.computer
Yes	3+4 = 7	1	youth	high	no	fair	no
		2	youth	high	no	excellent	no
		3	middle age	high	no	fair	yes
		4	senior	medium	no	fair	yes
No	1+ 2 = 3	5	senior	low	yes	fair	yes
		6	senior	low	yes	excellent	no
		7	middle age	low	yes	excellent	yes
		8	youth	medium	no	fair	no
		9	youth	low	yes	fair	yes
		10	senior	medium	yes	fair	yes
		11	youth	medium	yes	excellent	yes
		12	middle age	medium	no	excellent	yes
		13	middle age	high	yes	fair	yes
		14	senior	medium	no	excellent	no

For the low and medium, for the class variable buys computer, we are going to see how many yes is there. By looking at together, the level medium and low, there are 7 yes is there, and there are 3 no is there.

(Refer Slide Time: 15:28)

Tuples in partition D2

- High : D_2

High	Class: buys computer
Yes	2
No	2

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle age	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle age	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle age	medium	no	excellent	yes
13	middle age	high	yes	fair	yes
14	senior	medium	no	excellent	no

For high, because that was group D1, this is for group D2, how many yes is there, 2, how many no is there, 2 no is there. So, two yes is there and 2 no is there.

(Refer Slide Time: 15:38)

Gini index for income attribute

- The Gini index value computed based on this partitioning is

$$\begin{aligned}
 & Gini_{income \in \{low, medium\}}(D) \\
 &= \frac{10}{14} Gini(D_1) + \frac{4}{14} Gini(D_2) \\
 &= (10/14) (1 - (7/10)^2 - (3/10)^2) + \\
 &\quad (4/14) (1 - (2/4)^2 - (2/4)^2) \\
 &= 0.443 = Gini_{income \in \{high\}}
 \end{aligned}$$

The Gini index for income attribute. The Gini index value computed based on this partitioning is Gini income to the set low and medium, so $(10 / 14)$ Gini D1 + $(4 / 14)$ Gini D2, so how we got this four, by looking at the low and medium, there are 10 values is there, out of 14, so $10 / 14$. The another set, the remaining is 4. In D2, there is only 4, so $4 / 14$. So Gini D we have seen in the previous lecture, $1 - (7 / 10)$ whole square – $(3 / 10)$ the whole square.

How we got this 7, you see that there are 7 yes and there are 3 no. That is why it is $7 / 10$ whole square minus $3 / 10$ whole square. For D2, there are two yes, there are two no. So, $(4 /$

14) $(1 - (2 / 4) \text{ whole square} - (2 / 4) \text{ whole square})$, so this value is 0.443, so this is Gini value for the income high. For example, if you found the Gini value for low and medium, that is equivalent to finding the Gini value for the next group, that is the high.

(Refer Slide Time: 17:00)

Gini index for income attribute

- Consider the subset{high, medium}
- This would result in 10 tuples in partition D1 satisfying the condition "income \in {high, medium}"
- The remaining four tuples of D (low) would be assigned to partition D₂

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle.aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle.aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle.aged	medium	no	excellent	yes
13	middle.aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Consider the next subset, that is high and medium. When you look at high and medium, there will be 10 tuples in partitioning D1, so here also we are going for two group, one group is D1, and another group is D2. Here high and medium is in one set, that is D1, so when you go for high and medium, obviously the next will be low. Low will be in the set D2. High and medium, there will be 10 tuples satisfying the condition, remaining 4 tuples of D would be assigned to partitioning D2.

(Refer Slide Time: 17:57)

Tuples in partition D1

- High + Medium:

Medium + high	Class: buys computer
Yes	2+4
No	2+2

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle.aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle.aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle.aged	medium	no	excellent	yes
13	middle.aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

When you look at high and medium together, 6 yes for buying computer, 4 no for buying computer.

(Refer Slide Time: 18:10)

Tuples in partition D2

- Low :

Low	Class: buys computer	RID	age	income	student	credit.rating	Class: buys.computer
No	1	1	youth	high	no	fair	no
		2	youth	high	no	excellent	no
		3	middle.age	high	no	fair	yes
		4	senior	medium	no	fair	yes
Yes	3	5	senior	low	yes	fair	yes
		6	senior	low	yes	excellent	no
		7	middle.age	low	yes	excellent	yes
		8	youth	medium	no	fair	no
		9	youth	low	yes	fair	yes
		10	senior	medium	yes	fair	yes
		11	youth	medium	yes	excellent	yes
		12	middle.age	medium	no	excellent	yes
		13	middle.age	high	yes	fair	yes
		14	senior	medium	no	excellent	no

Then the next group, that is the tuples in partition D2, for low, there is one person has answered no, and 3 people has answered yes for buys a computer.

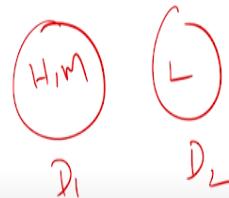
(Refer Slide Time: 18:22)

Gini index for income attribute

- The Gini index value computed based on this partitioning is

$$\text{Gini}_{\text{income} \in \{\text{high, medium}\}}$$

$$\begin{aligned}
 &= \frac{10}{14} \text{Gini}(D_1) + \frac{4}{14} \text{Gini}(D_2) \\
 &= (10/14) (1 - (6/10)^2 - (4/10)^2) + \\
 &\quad (4/14) (1 - (1/4)^2 - (3/4)^2) \\
 &= 0.45 = \text{Gini}_{\text{income} \in \{\text{low}\}}
 \end{aligned}$$



Now, we will find out Gini index value computed based on this partitioning. Gini index for income attribute, Gini index value based on the computed partitioning for Gini is for income, we had two category, one is D1 and D2. D1, we had high and medium, and D2 we had low. Now for high and medium, let us find out the Gini index because there was 10 out of 14 that comes high and medium, so the Gini for D1 plus 4, in this low, there was 4 out of 14, so Gini for D2.

So, what is the value of D1, because in D1, for high and medium together, there was 6 yes was there and 4 no was there. For D1, that is the Gini index value. For D2, the Gini index value is $4 / 14$, 1 yes was there and 3 no was there. So $1 / 4$ whole square minus $3/4$ whole square, so the Gini index for high and medium group is 0.45, that is nothing but the Gini index for the group also.

(Refer Slide Time: 19:54)

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle.aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle.aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle.aged	medium	no	excellent	yes
13	middle.aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

- Consider the subset{high, low}
- This would result in 8 tuples in partition D1 satisfying the condition "income \in {high, low}"
- The remaining six tuples of D (medium) would be assigned to partition D2

Now, we go for another subset, high and low. Now, here what we are going to do. We are going to have two groups because it is a binary classification, so high, low is one group, so this is D1, and D2 obviously it is D2. So here for this, this would result that is 8 tuples in partition D1 satisfying the income is high and low, the remaining 6 tuples of D would be assigned to partitioning D2.

(Refer Slide Time: 20:27)

Tuples in partition D2

- Medium:

Low	Class: buys computer
No	2
Yes	4

ID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle.age	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle.age	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle.age	medium	no	excellent	yes
13	middle.age	high	yes	fair	yes
14	senior	medium	no	excellent	no

For high and low, there was 5 yes is there, for high and low, there was 3 no is there. Another group is D2. In that, there was a medium. In that medium, there was a 2 no, and 4 yes.

(Refer Slide Time: 20:41)

Gini index for income attribute

- The Gini index value computed based on this partitioning is

$$\text{Gini}_{\text{income} \in \{\text{high, low}\}}$$

$$\begin{aligned}
 &= (8/14) (1 - (5/8)^2 - (3/8)^2) + \\
 &\quad (6/14) (1 - (2/6)^2 - (4/6)^2) \\
 &= 0.458 = \text{Gini}_{\text{income} \in \{\text{medium}\}}
 \end{aligned}$$



We will find out the Gini index value computed based on the partitioning that is for this group, high and low. In high and low, totally there was 8 value, 8 / 14, in that 1 – 5 yes out of 8, so 5/8 whole square, there are 3 no, -3 / 8 whole square plus for group D2, that is medium that was in another group, in that 6 elements was there out of 14, (6 / 14) (1 – how many yes was there, two no was there, 2 / 6 whole square, four yes was there, 4 / 6 whole square). So, the Gini index for the group high and low that is equivalent to. For the medium group, Gini index value, that also 0.458.

(Refer Slide Time: 21:40)

Gini Index values

	Gini Index values
Gini $\text{income} \in \{\text{high, low}\}$	0.458
Gini $\text{income} \in \{\text{high, medium}\}$	0.45
Gini $\text{income} \in \{\text{medium, low}\}$	0.443

We have completed all possible binary classifications, so Gini income high and low, Gini index value 0.458, the Gini group for high and medium is 0.45, for medium low, the Gini index value is 0.443. How to interpret this table, so this value 0.443, this is having the minimum Gini index value.

(Refer Slide Time: 22:11)

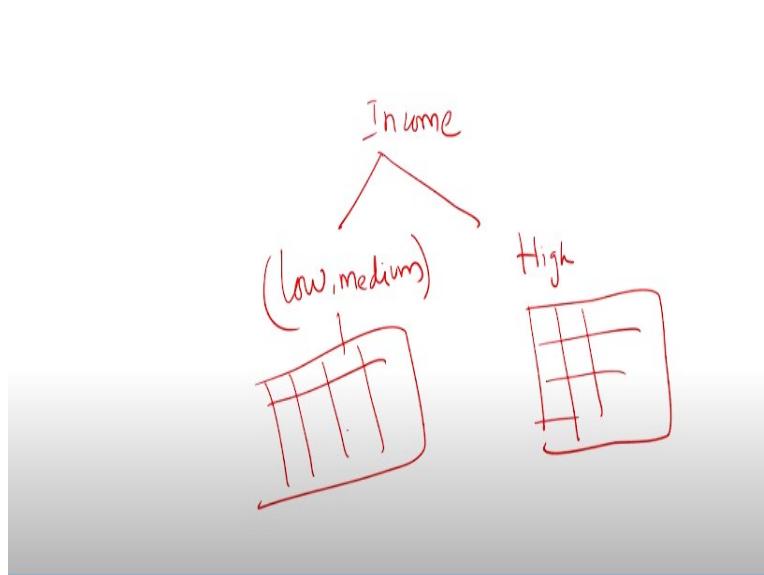
Interpretation

- The best binary split for attribute income is on $\{\text{medium, low}\}$ (or $\{\text{high}\}$) because it minimizes the Gini index
- The splitting subset $\{\text{medium, low}\}$ therefore give the minimum **Gini index for attribute income**
- Reduction in impurity = $0.459 - 0.443 = 0.016$**
- Further we calculate the same for the rest 3 criteria (age, student, credit rating)
- The one with minimum Gini index value will results in the maximum reduction in impurity of the tuples in D and is returned as the splitting criterion

The best binary split for the attribute income is on medium and low, otherwise high, because it minimizes the Gini index. When you look at the previous table, it is having the minimum Gini index value. The splitting subset, medium and low therefore gives the minimum Gini index for attribute income. So, the reduction in impurity is 0.459, this value which we got from slide number 18, this for the class D, - 0.443.

We got from this value, Gini index value for each subset. So the difference is 0.016. Further, we calculate the same for the rest of 3 criteria, so we got for reduction impurity for, this is income, there are another 3 attributes that is age, student and credit rating. For each attribute, we have to find out the reduction in impurity. The one with minimum Gini index value will result in the maximum reduction in impurity of the tuples in D and is returned as the splitting criterion.

(Refer Slide Time: 23:28)



Now, how to have the classifications. For example, we have income is the splitting variable. There was a 2 binary split. The first one was low and medium is one group, and high is another group. Now, in the high also, you might have some other table like this. For each value, we have to find out the Gini index, for low and medium group also, we will have another table. For this value also, we have to find out the Gini index.

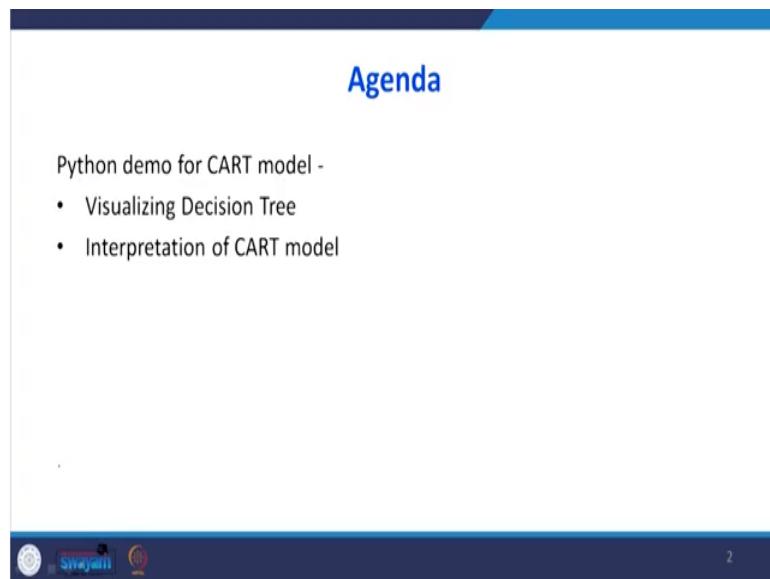
After finding out the Gini index, whichever is having highest level of reduction impurity, otherwise lowest value of Gini index should be chosen as the splitting criteria for further classification. In this lecture, I have explained how to choose an attribute for the decision tree model, by using two criteria. One is gain ratio and Gini index. For both the method, I have taken a numerical example. With the help of numerical example, I have explained how to choose an attribute. In the next lecture, we are going to use Python for making a CART model. Thank you.

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Computer Science and Engineering
Indian Institute of Technology - Roorkee

Lecture – 60
Classification and Regression Trees (CART) - III

In my previous lecture I have explained how to choose an attribute for the decision tree model. We know that there are three methods one is information gain method; second one is gain ratio method; third one is the Gini index. In previous lecture I have explained by using gain ratio and Gini index how to choose an attributes and I have explained all the procedures.

(Refer Slide Time: 00:49)



Agenda

Python demo for CART model -

- Visualizing Decision Tree
- Interpretation of CART model

In this lecture, we are going to use python with help of python. We are going to construct the CART model then I am going to explain the decision tree then I am going to interpret the output of CART model.

(Refer Slide Time: 01:07)

Example

Problem Description-

Han, J., Pei, J. and Kamber, M., 2011. *Data mining: concepts and techniques*. Elsevier.

<i>RID</i>	<i>age</i>	<i>income</i>	<i>student</i>	<i>credit_rating</i>	<i>Class: buys_computer</i>
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

3

This was the sample data this is an example say there are variable age, income, student, current rating is the attributes for independent variable the dependent variable is buys_computer. This example is taken from this source Han, Pei and Kamber the title of the book is Data Mining concepts and techniques.

(Refer Slide Time: 01:30)

Import Relevant Libraries and Loading Data File

```
In [1]: 1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt

In [2]: 1 data = pd.read_excel('CART.xlsx')

In [3]: 1 data
```

Out[3]:

<i>RID</i>	<i>age</i>	<i>income</i>	<i>student</i>	<i>credit_rating</i>	<i>buys_computer</i>	
0	1	youth	high	no	fair	no
1	2	youth	high	no	excellent	no
2	3	middle_aged	high	no	fair	yes
3	4	senior	medium	no	fair	yes
4	5	senior	low	yes	fair	yes
5	6	senior	low	yes	excellent	no
6	7	middle_aged	low	yes	excellent	yes
7	8	youth	medium	no	fair	no
8	9	youth	low	yes	fair	yes
9	10	senior	medium	yes	fair	yes
10	11	youth	medium	yes	excellent	yes
11	12	middle_aged	medium	no	excellent	yes
12	13	middle_aged	high	yes	fair	yes
13	14	senior	medium	no	excellent	no

4

I have brought this screenshot of first we will import relevant libraries and loading the data file. This file I have copied into the excel so I am importing pandas as pd import numpy as np, import matplotlib.pyplot as plt. I have imported the data then stored any object called data.

(Refer Slide Time: 01:50)

Methods used in Data Encoding

- **LabelEncoder ()**: This method is used to normalize labels. It can also be used to transform non-numerical labels to numerical labels.
- **Fit_transform ()**: This method is used for Fitting label encoder and return encoded labels.

Then there are different methods for encoding the data. The first one is LabelEncoder this method is used to normalize labels it can also be used to transform non-numerical labels into numerical labels. In our examples, the data are non-numerical that we are going to convert into numerical labels. The another function is fit under score transform this method is used for fitting label encoder and return encoded labels.

(Refer Slide Time: 02:23)

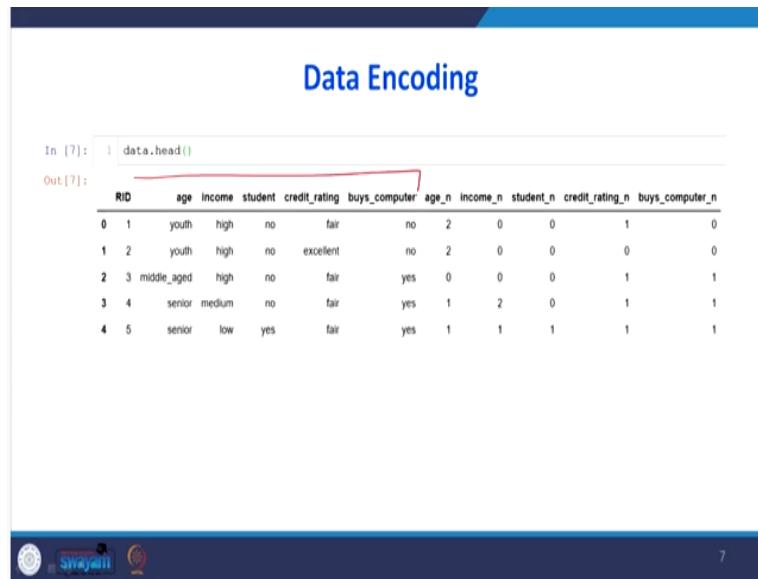
Data Encoding Procedure

```
In [4]: 1 import sklearn  
2 from sklearn.preprocessing import LabelEncoder  
  
In [5]: 1 le_age = LabelEncoder()  
2 le_income = LabelEncoder()  
3 le_student = LabelEncoder()  
4 le_credit_rating = LabelEncoder()  
5 le_buys_computer = LabelEncoder()  
  
In [6]: 1 data['age_n'] = le_age.fit_transform(data['age'])  
2 data['income_n'] = le_income.fit_transform(data['income'])  
3 data['student_n'] = le_student.fit_transform(data['student'])  
4 data['credit_rating_n'] = le_credit_rating.fit_transform(data['credit_rating'])  
5 data['buys_computer_n'] = le_credit_rating.fit_transform(data['buys_computer'])
```

So this is a data in coding percentages we import sklearn from sklearn dot preprocessing import LabelEncoder. So le_age LabelEncoder le_income like that for all the variables. Now there are different attributes like age but age, income, student credit rating buys computer this is in the text form. So that I am going to convert into numerical form, but new variable is age under n by

using this function le underscore age dot fit underscore transformation. So that wherever there is a fit_transformation, that is a text data is going to convert into numerical form.

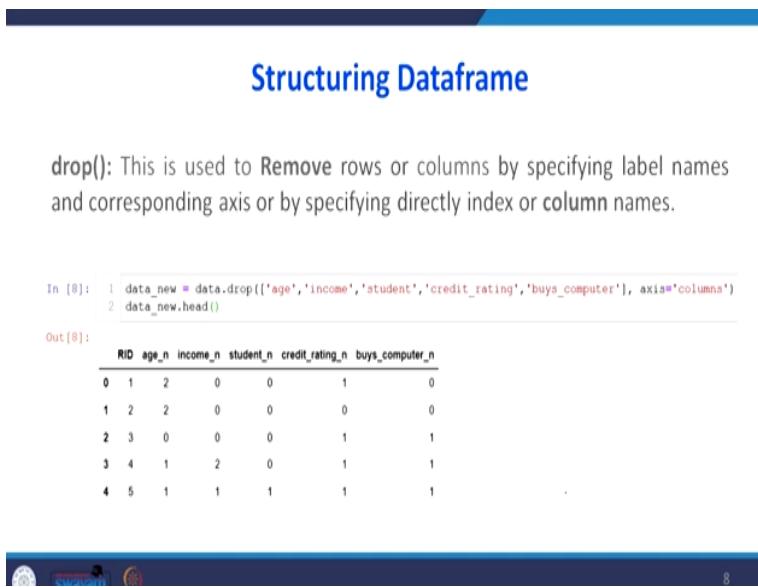
(Refer Slide Time: 03:10)



RID	age	income	student	credit_rating	buys_computer	age_n	income_n	student_n	credit_rating_n	buys_computer_n	
0	1	youth	high	no	fair	no	2	0	0	1	0
1	2	youth	high	no	excellent	no	2	0	0	0	0
2	3	middle_aged	high	no	fair	yes	0	0	0	1	1
3	4	senior	medium	no	fair	yes	1	2	0	1	1
4	5	senior	low	yes	fair	yes	1	1	1	1	1

Now what is happening is this portion is the text by using transformations I have converted into the numerical form.

(Refer Slide Time: 03:20)

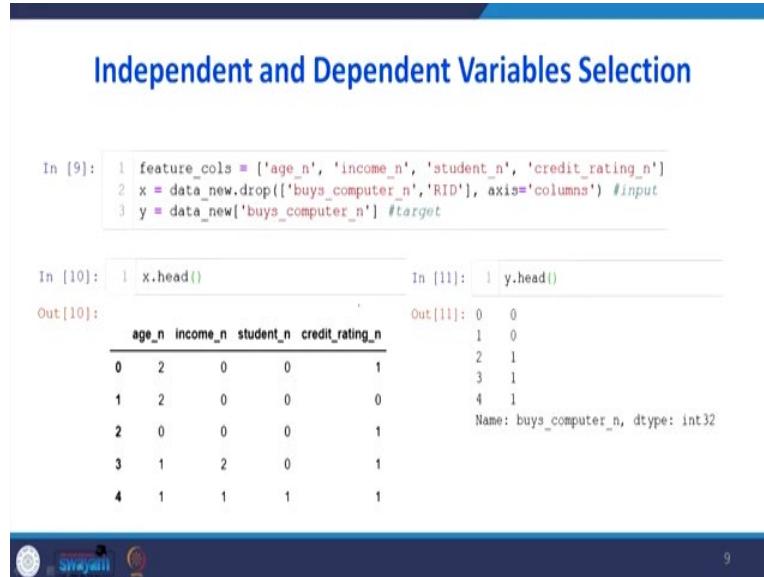


RID	age_n	income_n	student_n	credit_rating_n	buys_computer_n	
0	1	2	0	0	1	0
1	2	2	0	0	0	0
2	3	0	0	1	1	1
3	4	1	2	0	1	1
4	5	1	1	1	1	1

Then structuring the data frame by using the function drop. This is used to remove rows or columns by specifying label names in corresponding axis or by specifying directly index or column names. So what we are going to do in the previous layers, I told you there is a text data also is there , numerical data also there. Since already we are transforming into numerical that

text portions that I am going to drop it by using this drop function. So after that, this was the my dataset in this there is no text only numerical values. So this data set is going to be taken for building the CART model.

(Refer Slide Time: 03:58)



In [9]:

```
1 feature_cols = ['age_n', 'income_n', 'student_n', 'credit_rating_n']
2 x = data_new.drop(['buys_computer_n','RID'], axis='columns') #input
3 y = data_new['buys_computer_n'] #target
```

In [10]:

	age_n	income_n	student_n	credit_rating_n
0	2	0	0	1
1	2	0	0	0
2	0	0	0	1
3	1	2	0	1
4	1	1	1	1

In [11]:

	0	1
0	0	0
1	0	1
2	1	1
3	1	1
4	1	1

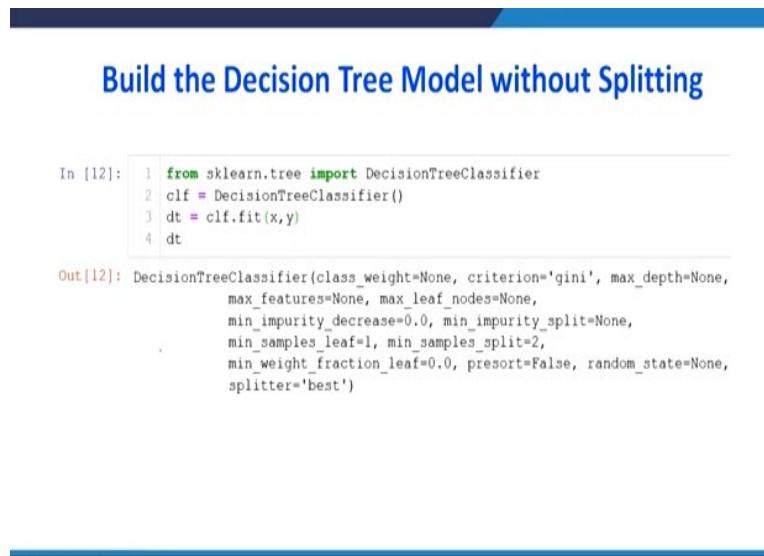
Out[11]:

```
0    0
1    0
2    1
3    1
4    1
Name: buys_computer_n, dtype: int32
```

9

In the building of the CART model we go to specify what is the independent and dependent variables we know that dependent variable is buys underscore computer. The independent variable is age, income, student credit rating you see that I am using age underscore n that is in the numerical form. In the dependent variables only two options, Yes or No that is buys underscore computer underscore n that has only two levels one is 0 or 1 that is Yes or No.

(Refer Slide Time: 04:31)



In [12]:

```
1 from sklearn.tree import DecisionTreeClassifier
2 clf = DecisionTreeClassifier()
3 dt = clf.fit(x,y)
4 dt
```

Out[12]:

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, presort=False, random_state=None,
splitter='best')
```

10

Now we are going to build the decision tree model without splitting what is the meaning of without splitting is in data mining whenever the huge amount of data is there, some data sets should be used for training the model the remaining data set should be used for testing the model. Now we are not going to do that way we are going to take all the dataset for build the model we are not going to test the model from sklearn.tree import DecisionTreeClassifier clf = DecisionTreeClassifier dt = clf.fit (xy) this was the dt this was the output.

(Refer Slide Time: 05:10)

The screenshot shows a Jupyter Notebook interface with a title bar 'Visualizing Decision Tree'. Below the title, there are three code cells:

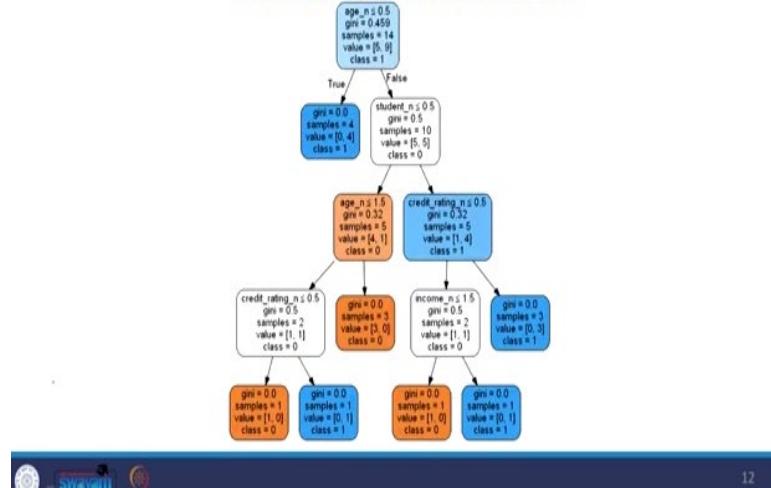
```
In [13]: 1 from sklearn.tree import export_graphviz  
2 from sklearn.externals.six import StringIO  
3 from IPython.display import Image  
4 import pydotplus  
  
In [14]: 1 dot_data = StringIO()  
2 export_graphviz(dt, out_file=dot_data,  
3                  filled=True, rounded=True,  
4                  special_characters=True, feature_names = feature_cols, class_names=['0','1'])  
5 graph = pydotplus.graph_from_dot_data(dot_data.getvalue())  
6 graph.write_png('buys_computer.png')  
  
Out[14]: True  
  
In [15]: 1 Image(graph.create_png())
```

At the bottom of the notebook, there are icons for file operations and a progress bar indicating the cell is running.

Now we are going to visualize the decision tree from sklearn dot tree import export graphviz. From sklearn dot externals dot six import String IO. From IPython dot display import Image then import pydotplus. So dot underscore data = StringIO export underscore graphviz this was the commands for getting the graphical output of our CART model. Then I have specified what is the dependent variable.

(Refer Slide Time: 05:48)

Decision Tree Visualization



12

This is the output of our CART model that is the Decision Tree this stage, let us understand how to interpret this. When age underscore n ≤ 0.5 there are two possibility it is true and false. I will explain the meaning of this 0.5 in the next slide. So whenever is a true, whenever this blue box represents it is a favorable decision for us. For example, orange colored box represents unfavorable.

What is unfavorable? That person will not buy the computer that is a class 0. If it is a class 1 means that person will buys the computer. Suppose if you want to interpret this blue box how to interpret is first look for the age, age we have seen, there are different levels in age. If it is true that person surely will buy the computer. If it is a false, then look for the student and the student also there is a two possibility Yes or No, this is true, this is false.

When you go for a student then this condition is failed it will go for a false then it will look for another attributes credit rating then it is true this is false. The credit rating when the false condition appears then this is favorable decision. If it is true, then look for another attribute income in that income if the false is applicable then it is favorable decision for us. So I will explain each values in the box and the meaning of the different color coding in coming slides.

(Refer Slide Time: 07:31)

Interpretation of the CART Output

Let us interpret the CART output.

(Refer Slide Time: 07:35)

Calculation of Gini(D)

- We first use the following Equation for Gini index to compute the impurity of D:

$$Gini(D) = 1 - \sum_{i=1}^m p_i^2$$
$$= Gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459.$$

ID	age	income	student	credit_rating	Class buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

14

This was the data which we have taken we first used the following equation for the Gini index to compute the impurity of D this also I have explained. We know that formula $Gini(D) = 1 - \sum_{i=1}^m p_i^2$, m is number of levels in our dependent variable here the m = 2 because Yes or No so Gini index = 1 minus the 9 represents how many years? 1, 2, 3, 4, 5, 6, 7, 8, 9; 9 yes out of 14 - this 5 represents number of Nos so 5 divided by 14 the whole square this is 0.459. This is Gini index for our dependent variable, so D represents the dependent variable.

(Refer Slide Time: 08:26)

Income Attribute

- Low, Medium, High
- Option 1: {Low, Medium}, {High}
- Option 2 : {High, Medium}, {low}
- Option 3 : {High, Low}, {Medium}

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle.aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle.aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle.aged	medium	no	excellent	yes
13	middle.aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

15

Now let us take one attribute at this stage suppose I have taken income, income is over attribute we are going to use Gini index for this attribute we know Gini index always go for binary classification. There are three levels low, medium, high option 1 is we can group into two way because it is a binary classification low medium is one group high is another group. Next high medium is one group low is another group the last choice is high, low is one group medium is another group for all these three combinations, let us find out what is the Gini index.

(Refer Slide Time: 09:11)

Tuples in partition D1

- Low + Medium:

Low + Medium	Class: buys computer	RID	age	income	student	credit.rating	Class: buys.computer
Yes	3+4 = 7	1	youth	high	no	fair	no
No	1+ 2 = 3	2	youth	high	no	excellent	no
		3	middle.aged	high	no	fair	yes
		4	senior	medium	no	fair	yes
		5	senior	low	yes	fair	yes
		6	senior	low	yes	excellent	no
		7	middle.aged	low	yes	excellent	yes
		8	youth	medium	no	fair	no
		9	youth	low	yes	fair	yes
		10	senior	medium	yes	fair	yes
		11	youth	medium	yes	excellent	yes
		12	middle.aged	medium	no	excellent	yes
		13	middle.aged	high	yes	fair	yes
		14	senior	medium	no	excellent	no

16

First, we will take low and medium in the low and medium we have to look at how many people answered Yes 1 No not this one 2, 3 that is how we got this 3 when it is medium, how many people answered Yes for our dependent variable, so this is 1 this is 2 this is 3 this is 4 so 3 + 4.

The same way when it is low, how many people answered No this case 1 only one options is there low and No.

Then the second case when it is medium, how many people answered No medium and No this is 1 then this, so 7 and 3 that is how we got this table this is one group part D1 actually what is happening here? so the income variable we are going to go for binary classification one is D1 another one is D2. D1 we are going to consider the two levels low and medium. In the another group we are going to consider only high so that is why D1 is low and medium so the D2 is only high.

(Refer Slide Time: 10:51)

Tuples in partition D2							
High	Class: buys computer	RID	age	income	student	credit.rating	Class: buys.computer
Yes	2	1	youth	high	no	fair	no
No	2	2	youth	high	no	excellent	no
		3	middle.ago	high	no	fair	yes
		4	senior	medium	no	fair	yes
		5	senior	low	yes	fair	yes
		6	senior	low	yes	excellent	no
		7	middle.ago	low	yes	excellent	yes
		8	youth	medium	no	fair	no
		9	youth	low	yes	fair	yes
		10	senior	medium	yes	fair	yes
		11	youth	medium	yes	excellent	yes
		12	middle.ago	medium	no	excellent	yes
		13	middle.ago	high	yes	fair	yes
		14	senior	medium	no	excellent	no

So D2 it is high how many people have answered when it is high, how many people answered Yes, this is 1 high Yes 2. So when it is a high, how many people answered No this 1 high No 2 yeah this is 2.

(Refer Slide Time: 11:13)

Gini index for income attribute

- The Gini index value computed based on this partitioning is

$\text{Gini}_{\text{income } \in \{\text{high, medium}\}}$

$$\begin{aligned}
 &= \frac{10}{14} \text{Gini}(D_1) + \frac{4}{14} \text{Gini}(D_2) \\
 &= (10/14) (1 - (6/10)^2 - (4/10)^2) + \\
 &\quad (4/14) (1 - (3/4)^2 - (1/4)^2) \\
 &= 0.45 = \text{Gini}_{\text{income } \in \{\text{low}\}}
 \end{aligned}$$

RID	age	income	student	credit.rating	Class: buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle aged	medium	yes	excellent	yes
13	middle aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

19



Now we are going to do the Gini index for income attributes the Gini index value computed based on this partitioning is, so Gini income belongs to low and medium so 10 to 12 / 14 how we got this 10 when you count low and medium, there will be 10 count it 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 divided by 14 Gini D1 the formula is $1 - (7/10)$ whole square – $(3/10)$ whole square. The second option for D2 4 / 14.

How we got 4? When it is a high? We have to count it how many high is there in the income. 1, 2, 3, 4 so $4/14$ Gini D2, D2 is 1 minus because number of Yes and number of Nos are same $1 - (2/4)$ whole square – $(2/4)$ whole square. So this is the Gini index for low and medium that is 0.443 that is equivalent to Gini index for other level high. The Gini index value for the another option what is the another option.

So there are two options, option one D1 D2 this is D1 this is D2, D1 high and medium is there D2 low is there. So when you having this two splits the Gini index for income attributes, which belongs to that is high and medium we are getting the 0.45.

(Refer Slide Time: 13:02)

Gini index for income attribute

- The Gini index value computed based on this partitioning is

$$\begin{aligned}
 \text{Gini}_{\text{income} \in \{\text{high, low}\}} &= (8/14) (1 - (5/8)^2 - (3/8)^2) + \\
 &\quad (6/14) (1 - (2/6)^2 - (4/6)^2) \\
 &= 0.458 = \text{Gini}_{\text{income} \in \{\text{medium}\}}
 \end{aligned}$$

ID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle aged	medium	no	excellent	yes
13	middle aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

20

The third option is high and low is one group then medium is another group. The same way when you continue, we are getting the Gini index is 0.4.

(Refer Slide Time: 13:18)

Gini index for income attribute

- $\text{Gini}_{\text{income} \in \{\text{low, medium}\}} = 0.443 = \text{Gini}_{\text{income} \in \{\text{high}\}}$
- $\text{Gini}_{\text{income} \in \{\text{high, medium}\}} = 0.45 = \text{Gini}_{\text{income} \in \{\text{low}\}}$
- $\text{Gini}_{\text{income} \in \{\text{high, low}\}} = 0.458 = \text{Gini}_{\text{income} \in \{\text{medium}\}}$

21

Now by comparing all the combinations, the lowest Gini index value is this one 0.443 that is where when income is high.

(Refer Slide Time: 13:31)

Gini index for Age attribute

- The Gini index value computed based on this partitioning is

$$\text{Gini}_{\text{Age} \in \{\text{Youth, middle_aged}\}} \\ = 0.457 = \text{Gini}_{\text{Age} \in \{\text{senior}\}}$$

$$\boxed{\text{Gini}_{\text{Age} \in \{\text{Youth, Senior}\}}} \\ = 0.357 = \text{Gini}_{\text{Age} \in \{\text{middle_aged}\}}$$

$$\text{Gini}_{\text{Age} \in \{\text{senior, middle_aged}\}} \\ = 0.393 = \text{Gini}_{\text{Age} \in \{\text{Youth}\}}$$

ID	age	income	student	credit_rating	Class: boys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

22

Now the Gini index for age attribute, so Gini index for senior it is 0.457 for middle age it is 0.357 for youth it is 0.393. The lowest Gini value is this one that is 0.357.

(Refer Slide Time: 13:51)

Gini index for student attribute

- The Gini index value computed based on this partitioning is

$$\text{Gini}_{\text{student} \in \{\text{Yes, No}\}}$$

$$= 7/14 (1 - (6/7)^2 - (1/7)^2) +$$

$$7/14 (1 - (3/7)^2 - (4/7)^2)$$

$$= 0.367$$

ID	age	income	student	credit_rating	Class: buys computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

23

Similarly, we will take another attribute that attribute is student, so for student attribute also when you find the Gini index it is 0.367.

(Refer Slide Time: 14:01)

Gini index for credit_rating attribute

- The Gini index value computed based on this partitioning is

$$\text{Gini}_{\text{credit rating} \in \{\text{fair, Excellent}\}}$$

$$= \frac{8}{14} (1 - (6/8)^2 - (2/8)^2) + \\ \frac{6}{14} (1 - (3/6)^2 - (3/6)^2) \\ = 0.428$$

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_agd	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_agd	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_agd	medium	no	excellent	yes
13	middle_agd	high	yes	fair	yes
14	senior	medium	no	excellent	no



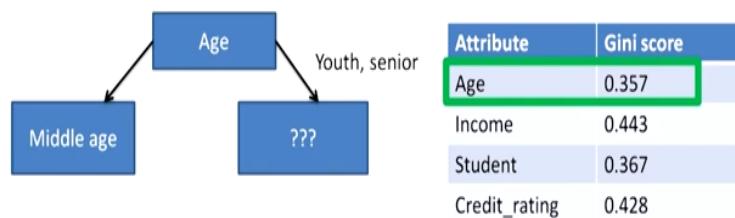
24

The next attribute is the credit_rating. For this, the Gini index is 0.428.

(Refer Slide Time: 14:09)

Choosing the root node

The attribute with minimum Gini score will be taken, i.e. Age ($\text{Gini}_{\text{Age} \in \{\text{Youth, Senior}\}} = 0.357 = \text{Gini}_{\text{Age} \in \{\text{middle_aged}\}}$)



25

Now when you bring all the generic index for different attributes, the age one is having the least Gini score that is 0.357 this attribute should be chosen for the classification that is the preference should be given for age. So the attribute with the minimum Gini score will be taken that is the age because that Gini index is 0.357. When you go for age that are 3 levels did middle age, youth and senior when you look at middle age, they have answered Yes for buying all people have answered Yes for buying computers with that we can stop it. Now there are youth and senior then we have to continue which attribute has to be chosen for here.

(Refer Slide Time: 14:54)

Gini index for different attributes for sample of 10

- After separating 4 samples belonging middle age, total 10 are remaining:

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

So since the all the middle aged people have answered Yes see that middle age is Yes, Yes, Yes. Now the next calculations we are going to drop these rows. After dropping these rows again here we are going to find out the Gini index, so after separating this 4 samples belonging to middle-age total 10 rows are remaining out of 14.

(Refer Slide Time: 15:22)

Gini index for different attributes for sample of 10

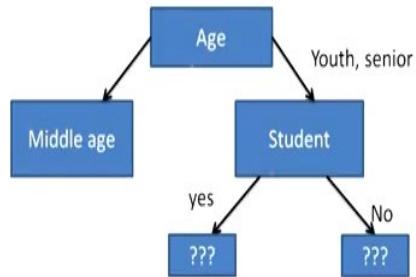
- Gini (D) = $(1 - (5/10)^2 - (5/10)^2) = 0.5$
- Gini_{Age} = 0.48
- Gini_{Credit Rating} = 0.41
- Gini_{Student} = 0.32
- Gini_{income} = 0.375
- Take student as node as it have mini. Gini Score

27

So for that 10 rows we are finding Gini index for our dependent variable 0.5 Gini index for age is 0.48 Gini index for our credit rating is 0.41 for a student it is 0.32 for income it is 0.375. Again, we have to look at which attribute is having the lowest Gini value. Take the student as node as it is having minimum Gini score.

(Refer Slide Time: 15:53)

Drawing cart



28

So what we have to do after first priority is age next, we have to take student as a classifier. In the student there was a 2 level was there Yes and No. If the student says Yes, then which attribute has to be chosen? If the student is No then which attribute has to be chosen, we will see that.

(Refer Slide Time: 16:10)

For branch Student = No

- Omit the marked rows
(Data entry), either
belonging Age =
middle_aged or student =
Yes
- Total 5 rows are remaining

RID	age	income	student	credit.rating	Class: buys.computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

29

Now what we have to do omit the marked rows either belonging age equal to middle-aged or student equal to Yes. So wherever we are talking about here in which variable which attribute has to be chosen. So what we are going to do, if it is a middle-age the dataset if the middle age is there that has to be dropped, the students are answering the student level is Yes that also is going to be dropped. So middle-aged is dropped wherever the student is Yes that also dropped. So how

many of you are going to drop 1, 2, 3, 4, 5, 6, 7, 8, 9 so out of 14, 9 rows we are going to drop it. So the remaining rows for further iteration is 5 rows.

(Refer Slide Time: 17:03)

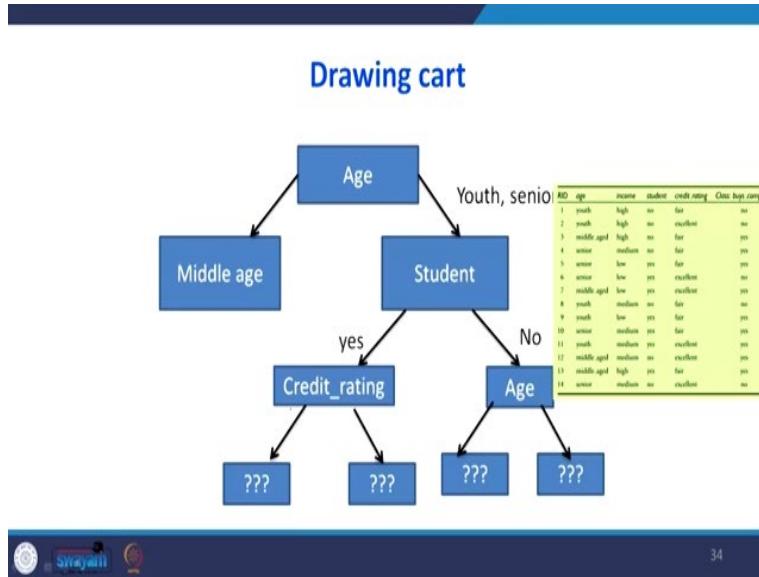
Gini index for different attributes For branch Student = No

- $\text{Gini}(D) = (1 - (4/5)^2 - (1/5)^2) = 0.32$
- $\text{Gini}_{\text{Age}} = 0.2$
- $\text{Gini}_{\text{Credit Rating}} = 0.267$
- $\text{Gini}_{\text{Student}} = 0.32$
- $\text{Gini}_{\text{Income}} = 0.267$
- Take age as node as it have mini. Gini Score

For that 5 rows again, we are going to find out the Gini index for the dependent variable 0.32 for the age 0.2 for credit rating, 0.267 for a student it is 0.32 for income it is 0.267. Again we have to look at which attribute is having the least Gini score. So the age is having the least Gini score. When the student level is No, we have identified what is next attribute. Similarly, when the student level is Yes, we have to find out what does the next year attribute for further classification for that omit the marked rows either belonging age equal to middle aged or student equal to No.

So when you do that one the remaining 5 rows are remaining using these 5 rows again, we are going to find out the Gini index for our dependent variable 0.32 for age 0.267 for credit rating it is 0.2 for student it is 0.32 for income it is 0.267. Again, you have to look at which attribute is having minimum Gini index though the credit rating is having the minimum Gini index.

(Refer Slide Time: 18:20)



What is to be done so the credit rating attribute has to be brought here for further classification.
Like this, we have to continue to satisfy all the conditions

(Refer Slide Time: 18:32)

Coding scheme

<u>Age</u>	Code	<u>Student</u>	Code
Youth	2	Yes	1
Middle Age	0	No	0
senior	1		

<u>Credit rating</u>	Code	<u>Income</u>	Code
Fair	1	High	0
Excellent	0	Low	1

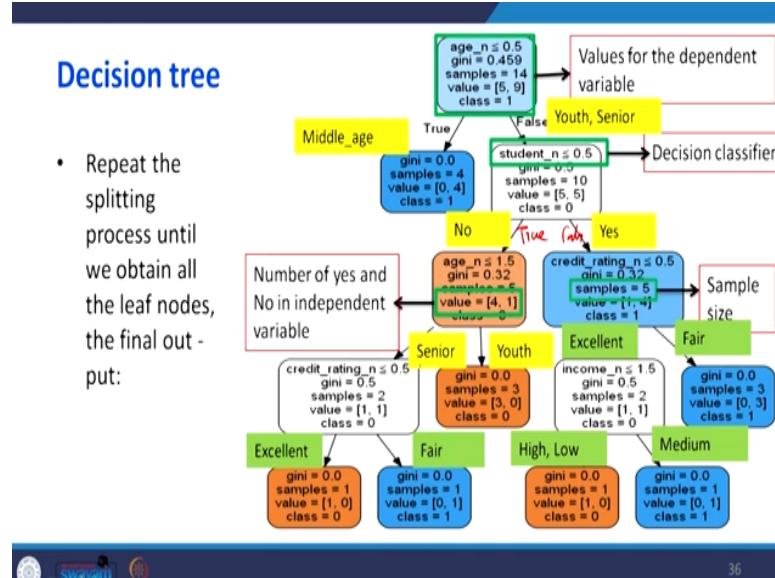
<u>Buys computer</u>	Class
Yes	1
No	0

Now I am going to explain the coding scheme because this coding scheme is important if you are going to interpret the python output for our CART model. There was a one attribute called age, student ,credit rating, income, Buys computer. So there are 4 independent variable, one dependent variable youth is coded as 2, middle-age is 0, senior is 1. So because this coding is important for interpreting the output.

(Refer Slide Time: 19:00)

Decision tree

- Repeat the splitting process until we obtain all the leaf nodes, the final output:



This was our python output Now if we look at the first variable is age underscore n when it is less than 0.5 this Gini index we got it 0.459 when you go back and see that this one, whatever the value which are appearing in the python output we have manually solved it so that you will feel more confidence when the age underscore n is the less than 0.5. Now you have to see what is the meaning of this 0.5 because we look at this coding the age is coded, youth is 2, middle-age is 0, senior is 1.

So if $n \leq 0.5$ that represents the middle age that is why we got this one when this condition is true this is middle age when the condition is false, they will belongs to youth and senior what is the meaning of this 5,9 the 5 represents No 9 represents Yes. Since in the middle aged group all are see that 0, 4 all are answered Yes. So the first represents No second represents Yes all are answered Yes so, we are not going for further classification.

Then, we are taking student underscore n as a another attribute for further classification when it is less than 0.5 we have to see what is less than 0.5 the student we have coded Yes = 1, No = 0, when I say $n < 0.5$ that represents No when the student is No that is why when the condition is true it is No this is true, this is false this is one branch when it is Yes there is another branch. Then next we choose age underscore n ≤ 1.5 .

Because now in the age only two group is there one is youth and senior when you go here youth and senior is there they when $n \leq 1.5$ that represents the senior when the condition is true that represents senior when the condition is false that represents youth we will go to the next one when the credit rating underscore $n < 0.5$. There are two options when the condition is true we got excellent. How are we got this excellent? Go to credit rating coding fair represents 1 excellent represents 0.

So when the condition is less than 0.5 it is excellent when the condition is true it is excellent if the condition is false it is fair. Then you go for income underscore $n \leq 1.5$ when the condition is true high and low how it is let us go for this coding income high = 0, low = 1, medium = 2 if it is 1.5 if $n < 1.5$ that represents low and medium when the income underscore $n < 1.5$ what is the meaning you look at this table.

So when it is less than 1.5 this group, this is less than 1.5 high and low, so high and low is one group medium is another group. So as a manager, how to interpret this first, the classifier is the age if it is true them go for middle age, the middle aged people will have the positive response for buying the computer that is why 1. If this condition fails then go for the student if they answer Yes then look for another attribute credit rating if it is fair, there is favorable response if it is excellent, we should go for further classification.

When it is excellent the next classifier is income when the income if it is true, they belongs to high and low then they will not buy the computer when it is medium then they will buy the computer. When we look at the left hand side, when the student $n \leq 5$ it is No then we will go for next attribute age because the age already we have dropped middle age the remaining is youth and senior if it is true it is senior, if it is false it is youth.

So if it is senior then you have to look for credit rating if it is youth there is a not favorable response. The 4 represents number of Nos 1 represents number of Yes in our dependent variable. So and another thing is look at the wherever the class is 1 there is only number of Yes is there see here that blue one 0, 4 only Yes is there 1, 4 only Yes is there 0, 3 only Yes is there 0, 1 Yes is there here also 0, 1 Yes is there.

So the blue boxes are which will give you the favorite decision for us the orange boxes and look at this, see that the here there is no see 1, 0 only No is there here also 1, 0 only No is there here also 3, 0 only No is there. So the orange box represents it is not going to give a favorable decision. The white box represents that is intermediate in the sense we have to go for choosing some more attributes for further conclusion.

So what does 14 represents values for the dependent variable there are 14 then the sample represents the sample size and so on. So repeat the splitting process until we obtain all leaf nodes and the final output. The leaf node represents this one this is leaf nodes.

(Refer Slide Time: 25:07)

The screenshot shows a Jupyter Notebook interface. The title bar says "Splitting Dataset". The main area contains two code cells:

```
In [12]: from sklearn.model_selection import train_test_split
```

```
In [13]: x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25, random_state=42)
```

At the bottom, there are three icons: a play button, a name tag, and a refresh symbol. The page number "37" is visible in the bottom right corner.

Now we are done the data without splitting, but in the data mining generally we used to split the dataset. So we split the dataset so testing data set 25% data is going to be used for testing the remaining data set is for the training.

(Refer Slide Time: 25:22)

Evaluating the Model

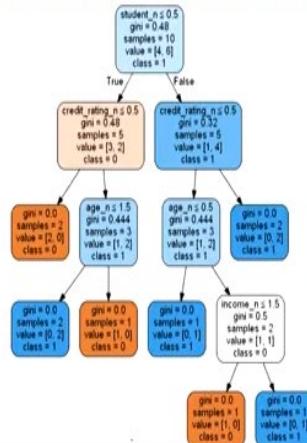
```
In [16]: 1 from sklearn import metrics  
  
In [17]: 1 y_pred = clf.predict(x_test)  
  
In [18]: 1 print("Accuracy:",metrics.accuracy_score(y_test, y_pred))  
Accuracy: 0.75  
  
True: [1 1 0 1]  
pred: [1 0 0 1]
```

39

So you run the python code, what I told previously. So here we are going to get the accuracy is 0.75 what is the meaning of this 0.75 our classifier, this decision tree model able to classify whether they are going to buy the computer or not with the 75% of accuracy. Then visualize the decision tree.

(Refer Slide Time: 25:46)

Decision Tree Visualization



41

So what is happening? you see when you are splitting the data set now the node variable is changed. The node is student previously it was the age, so what is happening since our data set is very only 14 dataset, we are getting this different result.

(Refer Slide Time: 26:03)

The screenshot shows a Jupyter Notebook interface with a data frame titled 'data'. The columns are RID, age, income, student, credit_rating, and buys_computer. The data consists of 15 rows with the following values:

RID	age	income	student	credit_rating	buys_computer
0	youth	high	no	fair	no
1	youth	high	no	excellent	no
2	middle_aged	high	no	fair	yes
3	senior	medium	no	fair	yes
4	senior	low	yes	fair	yes
5	senior	low	yes	excellent	no
6	middle_aged	low	yes	excellent	yes
7	youth	medium	no	fair	no
8	youth	low	yes	fair	yes
9	senior	medium	yes	fair	yes
10	youth	medium	yes	excellent	yes
11	middle_aged	medium	no	excellent	yes
12	middle_aged	high	yes	fair	yes
13	senior	medium	no	excellent	no
14	senior	medium	yes	excellent	yes

Now I am going to explain the python code for running the problem which I have explained. First import the pandas as pd than other libraries like numpy, matplotlib I have input. Then I am going to import the data. The data you know that that are age income, student ,credit rating, buys underscore computer. So this is in the text form, but I have to encode this data because they want it the numerical form I am encoding.

(Refer Slide Time: 26:46)

```

In [7]: data['student_n'] = le_student.fit_transform(data['student'])
In [7]: data['credit_rating_n'] = le_credit_rating.fit_transform(data['credit_rating'])
In [7]: data['buys_computer_n'] = le_credit_rating.fit_transform(data['buys_computer'])

In [7]: data.head()

Out[7]:
   RID    age  income  student  credit_rating  buys_computer
0     1  youth     high       no        fair         no
1     2  youth     high       no      excellent        no
2     3  middle_aged  high       no        fair        yes
3     4   senior  medium       no        fair        yes
4     5   senior     low      yes        fair        yes
5     6   senior     low      yes      excellent        no
6     7  middle_aged  low      yes      excellent        yes
7     8  youth  medium       no        fair         no
8     9  youth     low      yes        fair        yes
9    10   senior  medium      yes        fair        yes
10   11  youth  medium      yes      excellent        yes
11   12  middle_aged  medium       no      excellent        yes
12   13  middle_aged  high      yes        fair        yes
13   14   senior  medium       no      excellent         no

In [ ]: data_new = data.drop(['age','income','student','credit_rating','buys_computer'], axis='columns')
In [ ]: data_new
In [ ]: feature_cols = ['age_n', 'income_n', 'student_n', 'credit_rating_n']
In [ ]: x = data_new.drop(['buys_computer_n','RID'], axis='columns') #input
In [ ]: y = data_new['buys_computer_n'] #target

```

Now this shows that after encoding the right hand side we are able to see the equivalent numerical values. Now we are going to drop this text values because we are going to use only the numerical values for that building the CART model. So this was only the numerical values

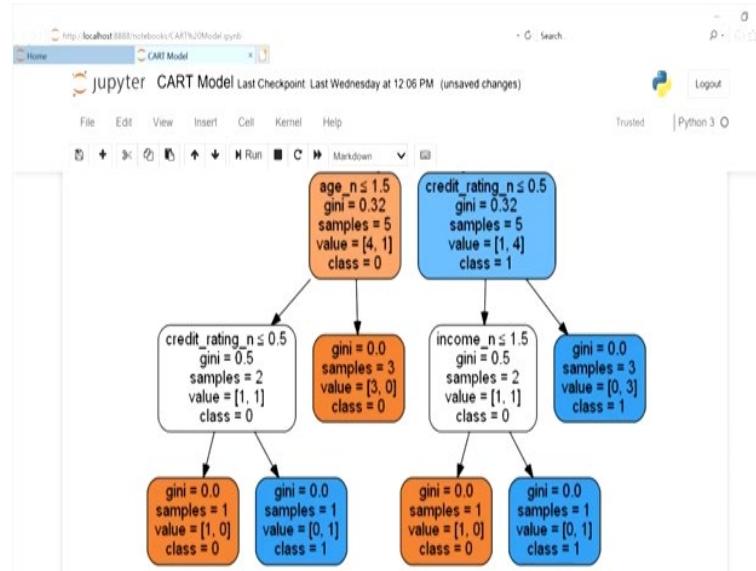
because there is a 14 dataset. Now we are going to declare what our independent variable, what are the dependent variable.,

(Refer Slide Time: 27:14)

	age_n	income_n	student_n	credit_rating_n
0	2	0	0	1
1	2	0	0	0
2	0	0	0	1
3	1	2	0	1
4	1	1	1	1
5	1	1	1	0
6	0	1	1	0
7	2	2	0	1
8	2	1	1	1
9	1	2	1	1
10	2	2	1	0
11	0	2	0	0
12	0	0	1	1
13	1	2	0	0

So the x these are the independent variable what are they? independent variable is there age, income, student, credit rating. Let us see what is the dependent variable? Dependent variable is buys underscore computer that has two levels, 0 and 1 then I am building the decision tree model for getting this output you need to install these packages.

(Refer Slide Time: 27:54)



This shows our CART output, so age is the first attribute and the age is true then they we are getting the leaf node it is a false then we are choosing another two attributes if it is then age,

credit rating, further we say credit rating then we go for income if it is false, we are stopping. So here what do you need to understand the blue, the blue circle represents, the blue rectangle represents the favorable decision for us, the orange one represents the unfavorable, white one represents in between that means we need to do further analysis. Now we are going to split the dataset splitting the ratio of 75, 25 so 75% days of data set is for training remaining 25% is for testing.

(Refer Slide Time: 28:59)

```

In [19]: | DecisionTreeClassifier?
Evaluating Model

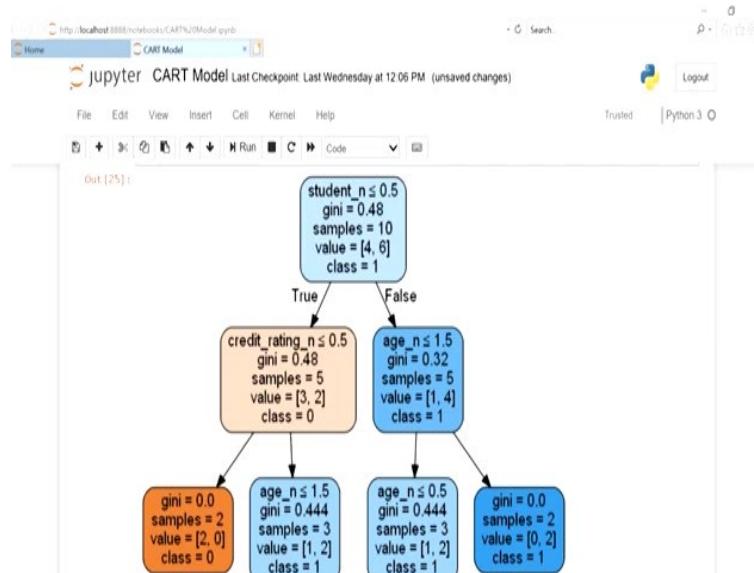
In [20]: | from sklearn import metrics
In [21]: | y_pred = clf.predict(x_test)
In [22]: | print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
Accuracy: 0.75

Visualizing Decision Tree
In [23]: | from sklearn.tree import export_graphviz
| from six import StringIO
| from IPython.display import Image
| import pydotplus

```

So after splitting the dataset, we are running the CART model then we are going to evaluate there is accuracy of our model. So the accuracy is 0.75 now we will visualize the CART model.

(Refer Slide Time: 29:40)



This was the output of data set where we are doing the splitting. So now the student is taken as the primary node for splitting if it is true, we will go for credit rating if it is false then you go for age for the next classifier. Then age we will go for the condition and this age underscore n ≤ 1.5 . I explained what is the meaning of 1.5 everywhere there are more possibility of favorable decision because class = 1.

There is orange rectangles which represents 0, 0 which is not favorable decisions. So what it means that when we split the data set, our decision making become very simple because our tree is in the very simple form, easy to interpret it. In this lecture I have explained how to do the CART model with the help of python. I have taken an example problem with the help of sample problem first I have got the CART model without splitting the dataset.

After that, after splitting the dataset, then I got output then I have compared, and I have explained in detail the output of the CART model. With that we are concluding this course data analytics with the python. Thank you very much for attending this course. Thank you.



**THIS BOOK IS NOT FOR SALE
NOR COMMERCIAL USE**



(044) 2257 5905/08

|



nptel.ac.in

|



swayam.gov.in