



Statistics

For Data Science I

WEEK 1 - WEEK 4

Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology Madras



IIT Madras
ONLINE DEGREE

INDEX

S. No	Topic	Page No.
	Week 1	
Introduction	Statistics for Data Science 1 - Introduction	1
Course Overview	Statistics for Data Science 1 - Course overview	5
Lecture 1.1	Introduction and Types of data : Basic definitions	27
Lecture 1.2	Introduction and Types of data : Understanding data	36
Lecture 1.3	Introduction and Types of data : Classification of data	54
Lecture 1.4	Introduction and Types of data : Scales of measurement	62
Tutorial 1.1	Tutorial 1 - Google Spreadsheets Introduction	75
Tutorial 1.2	Tutorial 2 - Formatting Google Spreasheets	85
Tutorial 1.3	Tutorial 3 - Spreadsheet formulae	96
Tutorial 1.4	Tutorial 4 - Downloading and uploading spreadsheets	110
	Week 2	
Lecture 2.1	Describing Categorical Data : Frequency distributions	121
Lecture 2.2	Describing Categorical Data : Charts of categorical data	138
Lecture 2.3	Describing Categorical Data : Best practices while graphing data	154
Lecture 2.4	Describing Categorical Data : Best practices while graphing data	163
Lecture 2.5	Describing Categorical Data : Mode and Median	169
Tutorial 2.1	Tutorial 1 - Problems Charts and Tables	190
Tutorial 2.2	Tutorial 2 - Problems Misleading Graphs	194
Tutorial 2.3	Tutorial 3 - SUMIF in Google Sheets	197
Tutorial 2.4	Tutorial 4 - VLOOKUP in Google Sheets	205
	Week 3	
Lecture 3.1	Describing Numerical Data : Frequency Tables for numerical data	220
Lecture 3.2	Describing Numerical Data : Mean	243
Lecture 3.3	Describing Numerical Data : Median and Mode	260
Lecture 3.4	Describing Numerical Data - Measures of dispersion- Range	276
Lecture 3.5	Describing Numerical Data - Percentiles, Quartiles, and Interquartile range	296
Tutorial 3.1	Tutorial 3.1	313
Tutorial 3.2	Tutorial 3.2	316
Tutorial 3.3	Tutorial 3.3	318
Tutorial 3.4	Tutorial 3.4	320
Tutorial 3.5	Tutorial 3.5	324

Tutorial 3.6	Tutorial 3.6	328
Tutorial 3.7	Tutorial 3.7	329
Tutorial 3.8	Box plot tutorial	332

Week 4

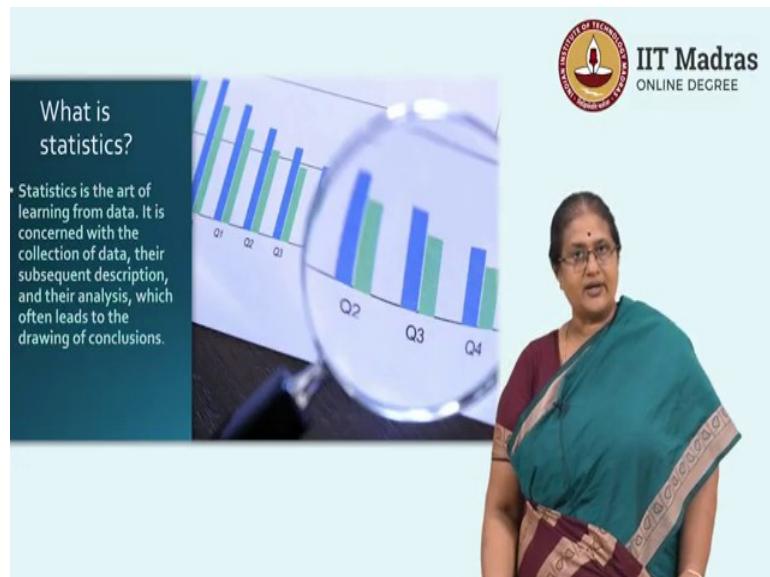
Lecture 4.1	Association between two variables - Review of course	335
Lecture 4.2	Association between two categorical variables - Introduction	338
	Association between two categorical variables - Relative	
Lecture 4.3	frequencies	352
Lecture 4.4	Association between two numerical variables - Scatterplot	373
	Association between two numerical variables - Describing	
Lecture 4.5	association	385
Lecture 4.6	Association between two numerical variables - Covariance	394
Lecture 4.7	Association between two numerical variables - Correlation	417
Lecture 4.8	Association between two numerical variables - Fitting a line	438
Lecture 4.9	Association between categorical and numerical variables	451
Tutorial 4.1	Tutorial 4.1	456
Tutorial 4.2	Tutorial 4.2	470
Tutorial 4.3	Tutorial 4.3	476
Tutorial 4.4	Tutorial 4.4	478

Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 01
Introduction

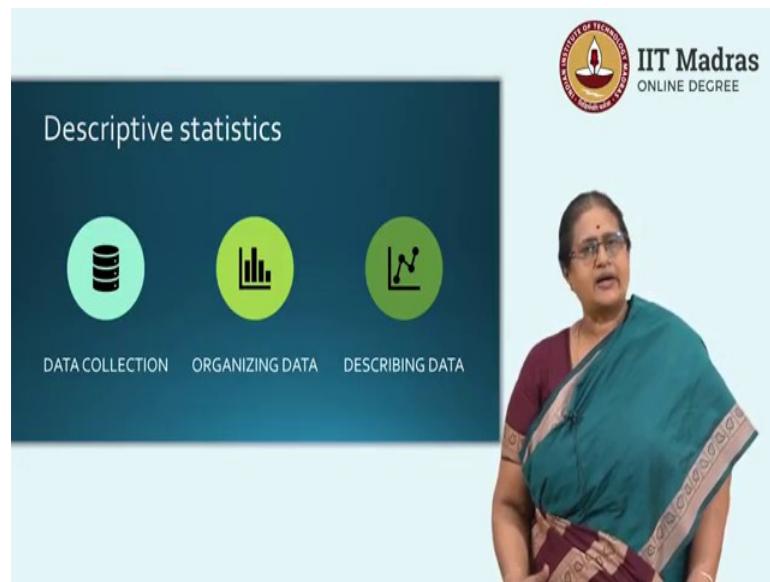
Hello. Welcome to Statistics for Data Science-1. I am Usha Mohan, Professor at the Department of Management Studies, IIT Madras. Over the next 12 weeks, we would travel together to understand the basics of statistics that is needed for your online degree programme. So, we first begin with understanding what is statistics.

(Refer Slide Time: 00:41)



So, first we look at what is the definition of statistics. So, one of the most popular definitions of statistics tells us statistics is the art of learning from data which includes collection of data, analysis of data, presentation of data, and then, drawing conclusions from data. So, one of the main things you are going to learn in this course is to first, we are going to look at how you collect data So, that is one of the main things which we will do.

(Refer Slide Time: 01:16)



So, we start with data collection. So, we look at how you collect data and then, we go on to see how you actually present this data. During this course, we learn about how you do present data in a tabular form; we introduce concepts of frequency tables; we introduce a concept of relative frequency and we look at how you tabulate data. We also learn about the different types of data that you encounter.

Then, we look at how you further look at describing data using graphical techniques. Some of the graphical techniques that you would learn are bar charts, pie charts, histograms, stem leaf plots etcetera. We then, also go about learning how you actually come up with numerical measures of data. What we look here mainly is at the measures of central tendency; namely, mean, median and mode and the measures of variation; namely, range, variance, standard deviation.

So, far we have just talked about describing a single variable, we also look at answering questions about association between variables. Towards this, we also introduce the concept of contingency tables, scatter plots and we also introduce what is a correlation matrix. So, predominantly in the descriptive statistics module, you will be learning about collecting data, organising data and describing data. Describing data both using graphical techniques and numerical techniques.

Now, once you are comfortable with just describing data, where the question is you just want to describe whatever data you have. The next step is to try and see I said that

statistics is the art of using data to draw conclusions. So, you want to infer something from data. So, this lays the playground for you to go next level which is inferential statistics. Inferential statistics will be taught in the next course on statistics. But, however, in this course, we will help you by laying the foundation or the building block to understand inferential statistics.

What do I mean by building block to understand inferential statistics? We motivate this through a small introduction to the probability theory.

(Refer Slide Time: 04:09)



When we go to probability theory, we start by what we mean by permutation and combinations. This is something which you are going to revisit. Most of you would have learned this in high school, but you are going to revisit the concept of permutation and combinations. Then, we introduce the notion of probability through random experiments like rolling a die, tossing a coin and then, afterwards we talk about probability of events, the combination of events and why they are necessary.

Finally, we end this course by introducing to two main distributions; namely, the binomial distribution and the normal distribution. Throughout the course, the focus is going to be on teaching at a conceptual level and applying these concepts to real world problems. So, you will have a lot of assignments which will test your knowledge, both at a conceptual level and an application level. There would be at every stage, we will try

and motivate the learner to look at understanding the concepts and applying their concepts through a set of case studies specifically designed for this course.

Welcome again to the world of statistics and this course which is Statistics for Data Science-1.

Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 02
Introduction - Course Overview

Hello. This is the Statistics for Data Science-1. This is the foundation course that you would require and this will give you the statistics. This is a preliminary course in statistics and in this lecture, we will be just going through and we will just see what the course overview is about; what are the learning objectives of this course and at the end of this course, what would you expect to learn.

(Refer Slide Time: 00:44)



Introduction and course overview

Week wise schedule and learning objectives



In this lecture, we start with a brief introduction to this course and we also will discuss as to what are the week wise schedule and learning objectives of this course. Now, why this course? What is the main learning objective of this course?

(Refer Slide Time: 01:00)



Statistics for Data Science-1 is an introductory course in statistics intended for beginners. Students learn to create handle data sets and summarize them using both graphical techniques and numerical techniques. Further, the notion of uncertainty is introduced and probability as a tool to handle uncertainty is discussed in detail. The concept of a random variable is introduced with a detailed discussion on the Binomial distribution and Normal distribution.



It is an introductory course. It is intended for beginners and by beginners, we mean by any person who has had tenth, class 10 level math. The, I think any person who has done math up to till; class 10 should be able to take this course comfortably. The main idea of this course is to help students learn to create data sets and summarize them and when we talk about summarizing them, we also talk about using both graphical techniques and numerical techniques.

The next important thing about this course is we are going to also introduce what we understand by the notion of uncertainty and the theory of probability not the mathematical theory. But we are going to use and discuss probability as a tool to handle this uncertainty. We will discuss about the using probability as a tool in some detail.

Finally, we introduce a notion or concept of a random variable and we focus on two important distributions; the binomial distribution and normal distribution and we will study applications about it. Throughout the course, we are going to focus only on applications and the understanding at a conceptual level. The focus is not going to be on the theory behind statistics and not behind theorems and proofs. The focus is going to be at an application level.

(Refer Slide Time: 02:37)



Course objectives

To provide students an understanding of statistics at a conceptual level to achieve the following objectives:

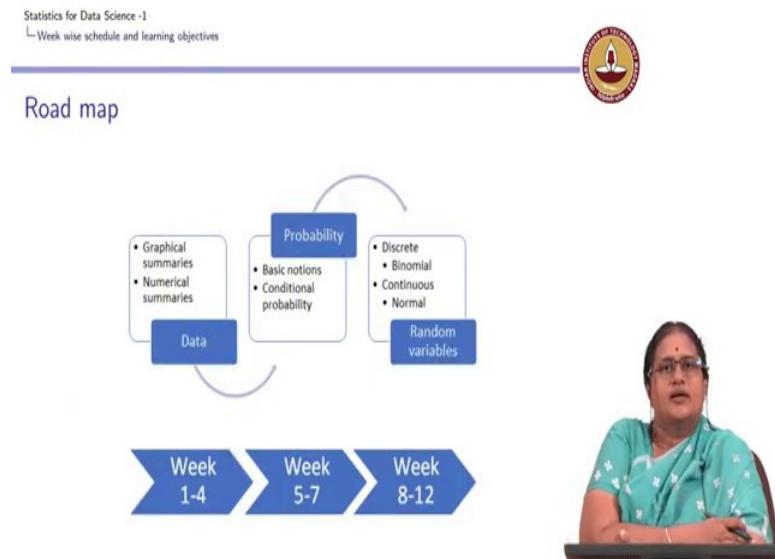
1. To create, download, and manipulate datasets.
2. To learn methods for presenting and describing sets of data.
Select an appropriate graphical technique for a given scenario.
3. To learn measures that can be used to summarize a data set.
Use of appropriate numerical summaries for a given scenario/question.
4. To understand uncertainty through probability.
 - 4.1 Understand notions of random experiment, events, probability and conditional probability.
 - 4.2 Understand use of random variables, both discrete (in particular, Binomial) and continuous (in particular, Normal).



So, what are the objectives of this course? The objective of this course is to provide students with an understanding at a conceptual level. At the end of the course, students should understand how to create download and manipulate data sets. This is one of the first objectives. The student should learn methods of presenting and describing data. By presenting and describing data, we expect the student would be able to understand what is an appropriate graphical technique; given a scenario, we talk about both graphical summaries of data and numerical summaries of data and very quickly, we will realize that numerical summaries cannot be given for all kinds of data. So, we would also focus on numerical summaries given a scenario and question.

So, all through the course, the focuses also going to be how students are going to formulate questions given data. We understand uncertainty through probability and as a course of understanding uncertainty, we are going to understand notions of what is a random experiment and understand use of random variables. So, this is at the end of this course, a student should be familiar about summarizing data, any kind of data and then afterwards have a good conceptual level of understanding of the basics of probability.

(Refer Slide Time: 04:07)



This course has been planned to span over 12 weeks and the roadmap for this course is given as in the following diagram; graph. So, if you modularize the course, we have 3 modules. In the first module, we are going to understand the basics of data; what are the types of data; how do we summarize data and we expect to achieve this in 4 weeks.

Once we know about data in this first 4 weeks, we are not dealing with any kind of uncertainty. But once we finish week 4, we move on to the concepts of probability, we learn basic notions of probability, we spend about 3 weeks to understand what are applications of probability and then, afterwards we move on to the notion of what we call random variables, wherein we talk about discrete random variables and continuous random variable with specific focus on the binomial and normal distribution.

(Refer Slide Time: 05:17)

Statistics for Data Science -1
└ Week wise schedule and learning objectives



Example

XYZ university has just completed admissions to their undergraduate program. Every admitted student fills up a form and the information is tabulated.

<https://docs.google.com/spreadsheets/d/15nJvZ-xBZDGb0oi-NCySIY4fETotXcJdm5pV1Fq2aI/edit?usp=sharing>

A portion of the data obtained by the admissions office is given below:

S.No	Name	Gender	Date of Birth	Marks in Class 10	Board	Marks in Class 12	Board	Mobile Number
1	Anjali	F	17-Feb-03	484	State Board	394	CBSE	xxx7252826
2	Pradeep	M	03-Jun-02	514	ICSE	437	ICSE	xxx5243748
3	Varsha	F	02-Mar-01	527	CBSE	442	CBSE	xxx5242824
4	Divya	F	22-Mar-03	397	State Board	401	Board	xxx6546889
5	Thomas	M	19-Dec-02	562	CBSE	451	CBSE	xxx4242736
6	Santa	F	19-May-02	533	ICSE	462	ICSE	xxx5242577
7	Prashant	M	30-Oct-01	496	CBSE	413	CBSE	xxx3352630
8	Ishaak	M	11-Feb-01	436	CBSE	375	CBSE	xxx1702736
9	Rafiq	M	31-Jul-02	501	ICSE	423	CBSE	xxx0029248
10	Bhavana	F	7-Apr-03	526	State Board	431	State Board	xxx5363036
11	Ashwani	M	25-Jan-00	450	State Board	398	CBSE	xxx7400862
12	Rohit	M	4-March-00	378	CBSE	291	CBSE	xxx4851749
13	Vikash	M	11-Oct-01	526	CBSE	436	ICSE	xxx2849482
14	Supriya	F	5-May-03	456	State Board	369	State Board	xxx300384
15	Nidhi	F	17-Nov-01	399	ICSE	400	ICSE	xxx5510065
16	Utkarsh	M	24-Jul-00	538	State Board	463	State Board	xxx8227401
17	Ayushman	M	19-Dec-0002	489	ICSE	402	ICSE	xxx5747800
18	Radhana	F	15-Aug-0001	529	CBSE	386	CBSE	xxx0069943
19	Anah	M	3-Jun-0003	420	ICSE	463	CBSE	xxx6254555
20	Mansi	F	7-Sep-0002	398	CBSE	388	ICSE	xxx0553867
21	Rahul Darshan	M	7-Aug-0001	510	State Board	390	State Board	xxx7361460
22	Nandini	F	24-July-0002	498	State Board	450	State Board	xxx8463927
23	Ishaak Thomas	M	20-Mar-0003	450	CBSE	425	CBSE	xxx0944647



So, what we are going to do now is going to start with what to expect from the course and give a week by week expectation, set the expectations for the course. Let us look at a very simple example, where I have a university let me call it a XYZ university which has just completed the admissions to their undergraduate program. Every admitted student has asked to fill up a form and the following information is tabulated.

(Refer Slide Time: 05:46)

S.No	Name	Gender	Date of Birth	Marks in Class 10	Board	Marks in Class 12	Board	Mobile Number
1	Anjali	F	17 Feb, 2003	484	State Board	394	CBSE	xxx7252826
2	Pradeep	M	3 Jun, 2001	514	ICSE	437	ICSE	xxx5243748
3	Varsha	F	2 Mar, 2001	527	CBSE	442	CBSE	xxx5242824
4	Divya	F	22 Mar, 2003	397	State Board	401	State Board	xxx6546889
5	Thomas	M	19 Dec, 2002	562	CBSE	451	CBSE	xxx4242736
6	Santa	F	19 May, 2002	533	ICSE	482	ICSE	xxx5242577
7	Prashant	M	30 Oct, 2001	496	CBSE	413	CBSE	xxx3352630
8	Harsha	M	11 Feb, 2001	436	CBSE	375	CBSE	xxx1702736
9	Rafiq	M	31 Jul, 2002	501	ICSE	423	CBSE	xxx0029248
10	Bhavana	F	7 Apr, 2003	526	State Board	431	State Board	xxx5363036
11	Ashwani	M	25 Jan, 2000	450	State Board	398	CBSE	xxx7400862
12	Rohit	M	4 March, 2000	378	CBSE	291	CBSE	xxx4851749
13	Vikash	M	11 Oct, 2001	526	CBSE	436	ICSE	xxx2849482
14	Supriya	F	5 May, 2003	456	State Board	369	State Board	xxx300384
15	Nidhi	F	17 Nov, 2001	399	ICSE	400	ICSE	xxx5510065
16	Utkarsh	M	24 Jul, 2000	538	State Board	463	State Board	xxx8227401
17	Ayushman	M	19 Dec, 2000	489	ICSE	402	ICSE	xxx5747800
18	Radhana	F	15 Aug, 2001	529	CBSE	386	CBSE	xxx0069943
19	Anah	M	3 Jun, 2003	420	ICSE	463	CBSE	xxx6254555
20	Mansi	F	7 Sep, 2002	398	CBSE	388	ICSE	xxx0553867
21	Rahul Darshan	M	7 Aug, 2001	510	State Board	390	State Board	xxx7361460
22	Nandini	F	24 July, 2002	498	State Board	450	State Board	xxx8463927
23	Ishaak Thomas	M	20 Mar, 2003	450	CBSE	425	CBSE	xxx0944647



So, the minute I say information is tabulated, you can see that. So, if you look at the data, this is just some hypothetical data we have created. Wherein, you can see that what the

information that was captured in the application form was the name of a person, the gender, date of birth, marks obtained in class 10, the board from which the student passed or wrote the exams; for class 10, marks obtained in class 12 and their board and their mobile number.

At the first look, you see that this is any all of us are familiar with data sets of this kind or data of this kind. So, what is the information that we normally would seek from this data set? So, what we are going to do is we are just going to start asking a few questions based on this data set. It is a very simple data set and we are trying to see that what are the type of answers that we can hope to answer during the course. So, given this data set, what we are going to do here? We are going to actually start you can see that this I just pasted a portion of the data set.

So, perhaps the first thing which we would want to see from this data set is to try and see that you already see that you have name, you have gender and you also see that when you are looking at the data set, you have some like marks in class 10 and marks in class 12 are captured as numbers. I do have mobile number also which is captured as numbers.

But when I talk about the board, I have both the state board and the CBSE, I have ICSE and I have date of birth is also given to me. So, the first thing which we notice when we see a data set of this kind is, we already know that I do not have one kind of variable. So, we need to understand what is the thing that is varying here.

(Refer Slide Time: 07:49)



1. Identify variables, observations



(Refer Slide Time: 07:53)

Week 1: Introduction



1. Understand how data are collected.
 - ▶ Identify variables and cases (observations) in a data set
2. Types of data- classify data as categorical(qualitative) or numerical(quantitative) data.
3. Understand cross-sectional versus time-series data.
4. Creating data sets; Downloading and manipulating data sets; working on subsets of data.
5. Framing questions that can be answered from data.



So, in week 1, we would introduce students to data sets of various kinds. So, the first objective in week 1 is to understand how data is collected. Once we understand how data is collected, we go back to understand what we mean by variable and observations. This is the first step in understanding data. So, before even jumping into doing the mechanics of statistics, we need to understand our data very well. When we need to understand our data very well, we need to understand what are the type of variables, what is the type of data we have collected and what are the different classifications that are available. The two major classifications which are available are we can classify data as quantitative and qualitative or numerical and categorical data.

So, the first thing, we are going to focus on is how do we create data sets or how do we download existing data sets. We might not be wanting to work with entire data set, we might be just wanting to work with a subset of data. So, how do we create the subsets of data? But the focus during this week 1 of our introduction would be on what are the questions that we are going to frame or what are the questions or what are the answers, we seek from data. Can we answer all these questions from data that we need to find out? But at least we want to train you in answering or in framing the right questions that you seek from data.

(Refer Slide Time: 09:39)



1. What is the gender diversity, in other words, what is the proportion of women students and proportion of male students?
2. How many students come from each board?



As an example, you can see that what we might want to ask from the data set that I presented to you earlier is we would want to know what is the proportion of women students to male students. This is a question which anybody would want to know or you would want to know what is the proportion of people who have come from CBSE? What is the number of people who have come from CBSE? What is the distribution of people who have come from ICSE or different boards? And if state of a student has been captured, we would like to know what is a regional representation of students?

So, there are many questions which you would want to know which are either we would like to know what is the proportion of people who are from a particular region or what is the count of people who are from a particular region. So, this is the type of questions pertaining to us, type of data which we refer to as categorical data.

(Refer Slide Time: 10:27)



Week 2: Describing categorical data- one variable

1. Organizing and graphing categorical data.
2. Create frequency tables for tabulated data.
3. Choosing an appropriate graphical technique for displaying data.
4. Discuss about misleading graphs.



So, the first-second week, we are going to spend time in understanding what is categorical data; what are the questions we want to find out from categorical data; what are the kind of answers, we seek from categorical data. And during this time, we are going to focus on how do we organize graph categorical data by creating frequency tables and the most important thing is though there are graphical techniques available to us, we are going to focus on which is an appropriate graphical technique to answer a particular question. This is the focus which we want to give is you frame a question and identify what is the right technique that you need to answer a question.

Many a time, we find that there are a lot of graphs which could be misleading. So, we would also focus on and discuss a bit of how important it is to correctly describe or correctly convey data through graphical summaries. Also, said that when you can broadly classify data into categorical data or qualitative data and numerical data. Once we understand how to summarize categorical data using graphical summaries, we move on to see what are numerical data.

(Refer Slide Time: 11:48)



1. What are the average marks obtained by students in Class 10/Class 12?
2. Is there a lot of variability in the marks obtained?
3. What is the least mark obtained? Highest marks obtained?
4. What is the average age of students admitted?



And what are the type of questions you have? The minute I say it is numerical data, we understand that we can do some arithmetic operation on it and get some mathematical summaries on it. One of the most often used summary is called the average or the mean. We are going to ask and what are the questions that, we hope to answer at this point of time. Again going back to a student data, one might be interested in knowing what was been the average marks obtained by the students in either their class 11, class 10 or class 12.

Another question people might want to know is has there a lot of variability in the marks that has been obtained or are people obtaining marks which are very close to each other. What is the least mark? What is the highest mark? What is the average age of students? So, you can see that these are very natural questions that many of us have been wanting to ask, when we are presented with a data set of that kind. In other words, the questions we seek to answer here involves some mathematical summary or a numerical summary.

(Refer Slide Time: 12:54)



Week 3: Describing numerical data- one variable

1. Visual representation of numerical data and interpret shape of distribution
2. Compute and interpret numerical summaries of data
 - 2.1 Compute and interpret measures of central tendency: mean, median, mode.
 - 2.2 Compute and interpret measures of dispersion: range, variance, standard deviation.
 - 2.3 Compute and interpret percentiles, Interquartile Range (IQR).
3. Compute and interpret five-number summary
4. Use histogram and box-plot to identify outliers in a dataset.



So, in week 3, we are going to go and focus on how you describe numerical data using numerical summaries. While we focus on numerical summaries at this point of time, we also discuss graphical summaries of continuous data; but the focus during this week is going to be pretty much on how you compute the well-known measures or how do you summarize this data using numerical measures. Broadly, measures of central tendency and variation.

So, we are going to focus on what are the main measures of central tendency and variation and we also will relate this to certain graphical things. Mainly, we are going to relate it to how you are going to have, how do you construct histograms and box plots. So, by the end of third week, you would know how to summarize a categorical data and numerical data; but till this point of time, we have been focused only on summarizing one variable.

(Refer Slide Time: 14:06)



1. Are there more women from state board when compared to men from state board?
2. Do students who have scored high marks in Class 10 score high marks in class 12 also?
3. Do students from State board score higher marks than those from other boards?



So, the next thing, what we would want to understand is given again go back to your school data set. In your school data set, you might want to ask questions like are women from state board, how do they; are there more women from state board compared to men from state board? Do people who have scored high marks in class 10? Do they also score high marks in class 12? So, we are asking questions now, where we are trying to understand whether two variables are related to each other or associated with each other.

I want to tell here a word of caution, we are not asking anything whether a causes b or b causes a. We are just interested in answering the question whether a student who obtained marks in class 10, does equally well in class 12. This is a very natural question to ask or do people from state boards score higher than marks than people from other boards. So, in a sense, we are looking at whether I can come up with measures or summaries which can actually summarize the relationship between variables. So, from moving from summarizing one variable, we move on to understand how do we capture association between variables.

(Refer Slide Time: 15:25)



Week 4: Association between two variables

1. Use of two-way contingency tables to understand association between two categorical variables.
2. Understand association between numerical variables through scatter plot; compute and interpret correlation.
3. Understand relationship between a categorical and numerical variable.



That is going to be the focus and the questions, we are going to ask in week 4 is to deal about association between two variables through what we call contingency tables, a graphical method which is called scatter plot, where we talk about numerical variables and we also understand how a categorical and numerical variable are related to each other. So, this is about the first module which is the week 1 to 4, where you would have had a reasonable and you should have a conceptual level understanding of what is a data set; what are the types of variables in my data set; how can I categorize them; how can I classify them as a quantitative, as a qualitative or as numerical or categorical and how do I summarize them.

When I talk about summaries, it is very important that we again, we ask a question is what is it I am seeking and what is the appropriate measure; be it a graphical summary or a numerical summary. I think the focus here is to be very clear about what is the question you are seeking to answer like the question you are seeking to answer, we are asking that what is the information you are seeking from the data set and how are you going to achieve it. This is the first module and at this end of the module this is where you should be. So, this is what you would be expected to know at the end of the fourth week.

(Refer Slide Time: 17:01)



After joining a college, the students want to form committees.

1. How many ways can a committee of 3 be formed from 10 people?
2. How many ways can a committee of 3 (President, Vice-president, and secretary) be formed from 10 people?
3. Basic principle of counting.



Now, the next module is the key module, where we are going to introduce the notion of probability. Now, why do we need to understand probability? Probability is extremely important because we live in uncertain times and what we want to know is we always ask questions, where there is an element of chance that is involved. Whenever there is an element of chance or whenever there is uncertainty, we need a very robust tool to handle this notion of uncertainty and probability is a good tool to handle this uncertainty.

But even before we understand what is the tool of probability, we need to understand something. We need to understand the basic principle of counting. Why do we need to understand the basic principle of counting? For example, after joining a student people; students typically after joining a college would want to form committees. The most natural questions you want to ask is how many ways a committee can be formed? How many ways can a committee of 3 can be formed? The difference between the first two questions is in first question, I was just interested in knowing the number of ways a committee of 3 can be formed; whereas, in the second question, I am interested in an order; I am interested in the President, Vice-president and secretary. Now, many of us would have been already introduced to this concept in high school which is famously known as permutations and combinations.

(Refer Slide Time: 18:33)



Week 5: Permutations and combinations

1. Understand the basic principle of counting.
2. Concept of factorials.
3. Understand differences between counting with order (permutation) and counting without regard to order (combination).
4. Use permutations and combinations to answer real life applications.



So, what we will do in week 5 is introduce a student to the basic principles of counting and we will understand how to apply the notion of permutation and combinations which is basically counting with order and counting without regard to order and the main focus of this basic principle of counting is to help a student understand how to use these permutations and combinations to answer real life applications. At the end of week 5, once you understand how this basic principle of counting is applied; then afterwards , we move on to said what is the; what are the questions we are really asking here.

(Refer Slide Time: 19:12)



Questions

1. What are the chances of a student getting a top grade?
2. What are the chances of a student getting a top grade given the student is from a particular board?
3. Key word is "chance"



So, when there is lot of uncertainty, a student has just joined college based on the marks, they have obtained in the 10 and 12th classes. So, the immediate thing, they would like to know is what is the chance of me getting a top grade? Now, the other questions you might want to ask is and from an administration, what is the chance of a student being a topper given or conditioned on the fact that they come from a particular region or they come from a they belong to a particular gender or they actually there lot of questions that you would want to ask, where you are actually asked the key word is chance; what is the chance.

(Refer Slide Time: 20:05)

Statistics for Data Science -1
└ Week wise schedule and learning objectives



Week 6-7: Probability

- 1. Understand uncertainty and concept of a random experiment.
- 2. Describe sample spaces, events of random experiments.
- 3. Understand the notion of simple event and compound events.
- 4. Basic laws of probability.
- 5. Calculate probabilities of events and use a tree diagram to compute probabilities.
- 6. Understand notion of conditional probability, i.e find the probability of an event given another event has occurred.
- 7. Distinguish between independent and dependent events.
- 8. Solve applications of probability.



So, we are used to this notion of a chance and this notion of a chance is basically captured and what we are going to introduce to the student to in the 2 weeks that follow is the basic notion of uncertainty. All of us know, you toss a coin, you know that there would be a head or a tail. But you really do not know whether the outcome is actually going to be a head or a tail. So, there is some uncertainty associated with this. We introduce the notion of what is randomness here and throughout these 2 weeks, what we are going to focus is we are going to focus on understanding what is a simple event or compound event.

And at this point of time, you should have learnt about sets in your math-1 course. Because probability and understanding the notion of simple events and compound events would need that we start representing events as sets, representation of events or sets

requires some idea of set algebra and at this point of time, you would have already had an introduction to set operations and set algebra from your math-1 course. So, you will be applying those concepts here to develop the notions of probability.

So, at the end of week 7, you should know the concepts or what are mutually exclusive events? What do I mean by independent events, that is does the fact that I get a head in the - I am tossing a coin twice - getting a head in the first toss, does that affect my outcome of the second toss or are they independent of each other? So, these are the notions which we are going to help understand, we are going to help develop the probability framework to answer these questions. So, this is what you are expected to know at the end of the 7th week.

(Refer Slide Time: 21:58)



Suppose one of the questions asked in the questionnaire asked students to report the number of siblings(sisters and brothers) they have.

1. What is the chance that a randomly selected student has 2 siblings?



Now, till this time, we have been always focused on events. I said we are going to talk about events as a set and we also know that when you talk about a set, it need not mean that you always have only numbers or which are elements of the set. But at some point of time, we need to ask questions. For example, in the same questionnaire which we refer to - suppose a student has been also asked to record or give the number of siblings; sisters or brothers they have and you are at by chance you are selecting some student.

A question you might want to answer is what is the chance that a randomly selected student from my database has 2 siblings? I am just restricting it to 2, but it could be 1, it

could be 0. In other words, I am associating a numerical value with whatever I want to achieve and this is what I am going to do through the concept of a random variable.

(Refer Slide Time: 23:06)



Week 8-9: Discrete random variables

1. Define what is a random variable.
2. Types of random variables: discrete and continuous.
3. Probability mass function, graph, and examples.
4. Cumulative distribution function, graphs, and examples.
5. Expectation and variance of a random variable.



So, we introduced the notion of a random variable at this point and we start with notion of discrete random variable, where the random variable takes discrete values numerical values and at this point of time, at the end of the random variable once we introduce what is a random variable, we would like to since it takes values describing some sort of a summary with this random variable would make sense. For example, you might want to know what are the on an average how many students have siblings, 2 siblings or what is the average number of siblings people have from the university.

So, these are the questions which you would want to answer. So, we introduce notion of expectation and variance towards answering these questions. Once we introduce a student to discrete random variables, we focus on a very important distribution; we spend 1 week to try and look at the questions for example.

(Refer Slide Time: 24:05)



Questions

A multiple-choice examination has 4 possible answers for each of 25 questions.:.

1. What is the chance of getting exactly 5 questions correct just by guessing?
2. What is the chance of getting more than 5 questions correct just by guessing?



We might want to know your all students attempt and I am sure that all of us are at some point of time, we have taken examinations which has multiple choice questions. And if you are certain, if you know the answer there is no uncertainty there; but many of us guess answers in a multiple choice question. So, the natural question to ask when you are writing a multiple examination which has multiple choice questions is what is the chance of me getting questions correct just by guessing. This is a very important question. This is a very natural question for us to know.

Again, guessing is an element of chance. So, here we are focusing on getting an answer right or wrong. So, experiments of this kind constitute what is called a Bernoulli experiment and then, after wards the distribution that helps us answer questions which are very similar to the type of questions which we are posing now, from what are called Binomial distribution.

(Refer Slide Time: 25:10)



Week 10: Binomial distribution

1. Understand the binomial distribution.
2. Applications of binomial distribution.



(Refer Slide Time: 25:16)



Questions

The time taken to write a test is recorded for each student. What is the chance that

1. the student requires more than 45 minutes to complete the test?
2. The student requires between 30 to 45 minutes to complete the test?



So, we are going to spend week 10 in understanding Binomial distribution and the focus again here is going to be on applications of binomial distribution. Now, till week 10, we have focused on random variables which take discrete values. But many a time, we are interested in answering questions, for example, if I am recording the time taken by a student to write an examination. The questions we are interested in knowing is what is the chance that the student requests more than 45 minutes to complete the test or what is the chance that the student would require between 30 and 45 minutes to complete the test?

So, the minute again you see that the question, we are answer asking is about the chance and the variable of interest here is time. So, we need a way to capture this variable of interest which is time and we notice that this time can take anything between 0 minute, 1 minute, one and half minutes, one and three-fourth minute. So, it is in a sense, it is a continuous variable. So, we focus on addressing this variable, these are called continuous variables; the type of questions we are going to ask on. So, first we will identify what are random variables that are continuous in nature.

(Refer Slide Time: 26:36)

Statistics for Data Science -I
└ Week wise schedule and learning objectives

Week 11-12: Continuous distributions and Normal Distribution

1. Concept of probability density function
2. The empirical rule of Normal distribution
3. Standard Normal distribution.
4. Applications of Normal distributions.

A woman in a teal sari is speaking on the right side of the slide.

And in the last week, we are going to focus on the last week 11 and 12; we are going to focus on variables that can take that are actually continuous in nature. Basically, we introduce the notion of probability density function; we are going to focus predominantly on a very very important distribution that arises again and again in our study of statistics which we refer to as the Normal distribution. We will focus on what is called the empirical rule of the normal distribution and again, the focus is going to be on applications of the normal distribution.

So, this has been the roadmap from week 1 to week 12. So, at the end of week 12, the student is expected to first differentiate between understand data, manipulate data sets, identify the types or classify the types of variables. To classify the type of variable, the student has to first know what is a variable; what is an observation and once you know that, how do I summarize these variables? To know how to summarize variables, a

student is expected again to ask questions, what are the why do I need to summarize a variable; what is the purpose to summarize a variable; what are the questions I need to answer; what is a questions I am seeking out of this data set?

What are the appropriate measures of summary; what are the appropriate summaries of the variable? Can I talk about association between variables? In the event of uncertainty, how do I handle uncertainty? We live in uncertain times, what is the basic notion of probability? What are the notion of random variables and what are the applications of these random variables? So, this is the course overview. This is what is expected. These are the learning objectives and at the end of the week 12, a student should be comfortable with the conceptual level understanding of whatever is presented so far.

Thank you.

Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 03
Introduction and Types of data – Part 1

In this week 1, the learning objectives are you understand first, why you are learning this course.

(Refer Slide Time: 00:21)

The screenshot shows a presentation slide with the title 'Statistics for Data Science -1' at the top left. Below the title is the Indian Institute of Technology Madras logo. The main content is titled 'Learning objectives' in blue. To the right of the text is a photograph of Prof. Usha Mohan, a woman with glasses wearing a blue and brown sari, standing behind a podium. At the bottom of the slide are standard presentation navigation icons.

Statistics for Data Science -1

Learning objectives

- 1. What is statistics?
 - ▶ Descriptive statistics, inferential statistics.
 - ▶ Distinguish between a sample and a population.
- 2. Understand how data are collected.
 - ▶ Identify variables and cases (observations) in a data set
- 3. Types of data-
 - ▶ classify data as categorical(qualitative) or numerical(quantitative) data.
 - ▶ Understand cross-sectional versus time-series data.
 - ▶ Measurement scales
- 4. Creating data sets; Downloading and manipulating data sets; working on subsets of data.
- 5. Framing questions that can be answered from data.

What is statistics? We are just going to tell briefly about the two main branches of statistics which will be relevant at this point of time to you people will tell what you mean what is understood by descriptive statistics and inferential statistics.

The minute I talk about inferential statistics, I need to introduce what is the notion of a sample and a population. So, that is what I am going to introduce. Then we move on to understand why we need data; we will understand a bit about how data is collected and we will talk about how to organize data in form of what we call a data set.

Once we have a data set, we will understand to more about data by classifying data in terms of categorical and numerical or cross-sectional and time-series, and we will talk a bit we will discuss a bit about measurement scales.

Finally, I think any statistical analysis, the key is to understand your data and frame questions based on data. So, we will focus some time to try and understand and train ourselves to frame questions based on data. So, these are the learning objectives for the week 1.

(Refer Slide Time: 01:36)

Statistics for Data Science -1
└ Introduction
└ Basic definitions

What is Statistics?



Definition

Statistics¹ is the art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis, which often leads to the drawing of conclusions.

¹Ross, Sheldon M. Introductory statistics. Academic Press, 2017.



What is statistics? If you go through the definitions of statistics over the years, you can see that there has been a transformation and that has been changing over the period of time. What started as just summarizing data, then afterwards gradually improve to inference from data and then afterwards now with lot of data available, statistics is being redefined as the art of learning from data.

Now, the minute I say learning from data, it includes that you want to seek some information from data. So, Sheldon Ross defined statistics as the art of learning from data, you are concerned with collection of data, subsequent description and their analysis which often leads to drawing of conclusion. So, the main idea of statistics and statistical analysis is to actually draw conclusions based on data.

(Refer Slide Time: 02:44)

Statistics for Data Science -1
└ Introduction
└ Basic definitions

Major branches of statistics

1. Description

Definition
*The part of statistics concerned with the description and summarization of data is called **descriptive statistics**.*

2. Inference

Definition
*The part of statistics concerned with the drawing of conclusions from data is called **inferential statistics**.*

► To be able to draw a conclusion from the data, we must take into account the possibility of chance- introduction to probability.



So, if you look at the classification of statistics, even though there are newer branches of statistics and new titles given, you may broadly classify the branches of statistics or you might broadly look at the main branches of statistics to be two: one way you are describing data that is a part of statistics which is concerned to description and summarization of data more popularly referred to as the descriptive statistics branch.

The part of statistics which is concerned with drawing conclusions from data is called the inferential statistics branch that is you want to infer from data. Now, when you want to infer from data, there is one very important thing which is the possibility of chance because when you are inferring from data there is an element of chance you do not have exactly what you are having what you know.

And, hence we are preparing in this course in this foundation course with an introduction to probability, to help you understand or help you prepare for the next league or the next course where will you where you will be learning about inferential statistics.

(Refer Slide Time: 04:11)

Statistics for Data Science -1
└ Introduction
└ Population and sample

Population and sample

POPULATION

SAMPLE

REPRESENTATIVE

Suppose we are interested in knowing

- ▶ The percentage of all students in India who have passed their Class 12 exams and study engineering.
- ▶ The prices of all houses in Tamil Nadu.
- ▶ The total sales of all cars in India in the year 2019.
- ▶ The age distribution of people who visit a city Mall in a particular month.



So, primarily when you talk about inferential statistics, we are trying to talk about drawing of conclusions from data. Now, a branch of inference as inferential statistics, one important thing is many a time you are interested perhaps in knowing about the percentage of all students in India who have passed their Class 12 exams and study engineering; the prices of all households in Tamil Nadu; the total sales of all cars in India in the year 2019; the age distribution of people who visit a city Mall in a particular month.

So, one way of answering all these questions is one is through a complete enumeration – you go and collect data on everybody or everything you are interested. For example, in this question you are interested in knowing about the percentage of all students in India, but very quickly you understand that getting this kind of data might not be very easy.

So, many a time what we are interested in knowing is the percentage of all students in India. Now, if I just want to construct a database and I would want the actual data of all the students who have passed class 12, but if my intention is just to know an overall feel of what are the kind of people who finally, end up taking engineering then one thing I would want to know is work with a smaller subset of all the students in India. All the set of all students in India is what we refer to as a population. A smaller subset of this is referred to as a sample. It is a subset, so, I am putting it as a sample.

Now, many a time you might be wanting to know about the prices of all houses. Again, you need not go and find out about all the houses that have been sold in a particular year; you might want to know about a smaller subset of the entire population. One thing you want about the sample is you want it to be as representative as possible you want the sample to be as representative as possible.

(Refer Slide Time: 06:42)

The slide is titled "Population and sample". At the top left, there is a navigation menu:
Statistics for Data Science -1
└ Introduction
└ Population and sample

On the right side of the slide is the logo of the Indian Institute of Technology (IIT) Madras.

The main content area contains two diagrams. The first diagram, labeled "Population", shows a large rectangular frame containing several colored bars (yellow and blue) of varying heights. The second diagram, labeled "sample", shows a smaller rectangular frame containing a subset of the bars from the population frame, also of varying heights.

Definition
The total collection of all the elements that we are interested in is called a *population*.

A video feed of a woman speaking is overlaid on the bottom right of the slide.

Now, what do we mean by representative sample? For example, let me define the population is a collection of all elements that we are interested in. If this is the population so, let me draw different colours here. What is the tool I use?

So, suppose this is a population and I take another subset here. Suppose I take a subset, this is a subset. The smaller set is actually a subset of the larger set, but we very quickly notice that the smaller set does not have any yellow elements in it. So, I cannot say this smaller set is actually a good representative sample of the larger set.

(Refer Slide Time: 07:39)

Population and sample



Definition

The total collection of all the elements that we are interested in is called a **population**.

Definition

A subgroup of the population that will be studied in detail is called a **sample**.



So, a sample is basically a subgroup of the population that will be studied in detail.

Now, we need the idea of a population and sample and you will be introduced to this concept of population and sample in greater detail when you do your inferential statistics course. But, nevertheless why do we need the concept of a population and sample in this course is, eventually when we are going to come up with summary statistics, we always need to understand whether the summary statistics is for a population or a sample and this is something which we will know in due course.

(Refer Slide Time: 08:15)

Purpose of statistical analysis



- ▶ If the purpose of the analysis² is to examine and explore information for its own intrinsic interest only, the study is descriptive.
- ▶ If the information is obtained from a sample of a population and the purpose of the study is to use that information to draw conclusions about the population, the study is inferential.
- ▶ A descriptive study may be performed either on a sample or on a population.
- ▶ When an inference is made about the population, based on information obtained from the sample, does the study become inferential.



So, what is the purpose of statistical analysis? Now, when would you use a descriptive statistics? When would you use inferential statistics? Now, if the purpose of your analysis is just to examine and explore information for its own intrinsic interest only, this study is descriptive.

Now, what do we mean by that? Let me demonstrate it to you through a data set ok.

(Refer Slide Time: 08:45)

S.No	Player Name	Jersey No	Matches played	Role	Runs	Batting Avg	Highest score	Wickets	Bowling Avg	Best bowling
1	Sachin Tendulkar	10	403	Batsman	18426	44.83	200	154	44.48	5/32
2	Virat Kohli	18	248	Batsman	11867	59.34	183	4	166.25	1/15
3	MS Dhoni	7	350	Batsman, WK	10773	50.58	183	1	31	1/14
4	Rohit Sharma	45	224	Batsman	9115	49.27	267	8	64.38	2/27
5	Sehwag	46	251	Batsman	8273	35.04	219	96	40.14	4/6
6	Gambhir	5	147	Batsman	5238	39.68	150	0	0	0/13
7	Yuvraj	12	304	All-rounder	8701	36.56	150	111	45.48	5/31
8	R Jadeja	8	165	All-rounder	2296	31.89	87	187	44.8	5/36
9	Zaheer Khan	34	200	Bowler	792	12	34	282	29.4	5/42
10	H Singh	3	236	Bowler	1237	13.3	49	269	33.38	5/31
11	Bumrah	93	64	Bowler	39	3.7	10	104	24.43	5/27
12	M Shami	11	77	Bowler	145	7.74	25	144	25.42	5/69
13	R Ahwin	99	111	Bowler	675	16.06	65	150	39.21	4/25
14	Kuldeep Yadav	23	60	Bowler	118	13.31	19	104	26.16	6/25
15	Y Chahal	6	42	Bowler	5	2.5	3	55	24.35	6/25
16	Hardik Pandya	33	54	All-rounder	951	29.91	83	54	40.64	3/31
17	Kedar Jadhav	81	73	All-rounder	1389	42.09	120	27	27.78	3/23
18	KD Karthik	21	94	Batsman, WK	1752	30.2	79	---	---	---
19	Robin Uthappa	6	46	Batsman	934	25.94	86	---	---	---
20	Ambati Rayudu	5	55	Batsman	1694	47.05	124	3	41.33	1/5
21	Rahul Dravid	19	344	Batsman	10889	39.16	153	4	42.5	2/43

This is again another hypothetical data set which is just showing the names of the cricket players. All of us are very well aware of these cricket players – Tendulkar, Kohli, Dhoni. The matches they have played, in what role, what are the total runs, the batting average, the highest score, wickets, bowling average and best bowling.

Now, suppose a purpose is just to understand what are the total runs scored, what is the batting average, what is the who has the highest batting average, who has the highest run scored, who have played the most number of matches, if these who has taken the highest number of wickets – if these are the questions of interest then all these questions of interest which I have just posed now, can directly be just got from the data set.

I might also want to order the number of runs of a batsman has scored; I might want to also know what is among the batsman how have the people scored runs and all of this I can just describe this data. I do not have to do anything more about this data. So, in this case the question I am asking is basically, the purpose I have here the purpose I have

here is to just examine and explore the information that is given. So, the study is just descriptive. I am not asking anything more. I just want to describe the data set that is given here and this study is descriptive.

But, suppose I am using this and one thing which we notice again in this data is the following. If you look at this data this data is not the entire cricketing data about all the cricketers available from all the countries.

It is a sample from an entire population of data. It is just a small sample. I can say it is at best a representative sample of the Indian cricketing data over the last 5 or 10 years or perhaps about this could be about for the in the last decade. It is a sample of definitely, it is a sample of the Indian cricketing data.

But, it is again not the entire population which includes over all batsmen and overall cricketers, but however, if I am just interested in summarizing this data if my inherent interest is just about summarizing this data, then I would be interested in only a descriptive nature of studies for which descriptive statistics is sufficient.

But, now if I am going to use this to draw a conclusions further conclusions; for example, if I want to know about the role a batsman plays with a batting average, I would need more information and I want to pick up a team for the future. For example, you we all know about the IPL auctions and how people are chosen. So, there is a further role. I am just not interested in describing this data.

The bigger role for me or the bigger interest for me is to use this data to gather or infer some information which I am going to use in my decision making process. For that I would I am going to have an element of chance and there I am going to have what I need is I am going to have an inferential study in that case. So, very often we see that a descriptive study we need to understand whether our nature of a study is only going to be descriptive or whether we want to do an inferential study.

When we come to inferential study a descriptive study sorry when we come to for a descriptive study it might be either performed on a sample or on a population. Since in the classes to come we will be talking about descriptive statistics in detail. We need to understand whether a descriptive study is performed on a sample or on an entire

population that is the reason why we introduced the notion of a sample and a population at this stage.

However, if our inference is to be made about a population based on the sample, then the study becomes inferential. Inferential statistics is not the scope of this course, but however, you will be introduced to the concept of probability which will help you develop the methodology towards inferential statistics.

(Refer Slide Time: 13:49)

The screenshot shows a presentation slide with the following elements:

- Navigation bar: Statistics for Data Science -1, Introduction, Population and sample.
- Section title: Summary.
- Image: Logo of the Institute of Technology, Roorkee.
- List: Descriptive statistics, Inferential statistics, Population and sample.
- Image: A woman in a blue sari sitting at a desk.
- Control icons: Back, forward, search, etc.

So, in summary, you should know the two main branches are descriptive statistics, inferential statistics. You are going to do a descriptive study or inferential study based on what is your purpose of study.

If your intrinsic purpose is just to summarize your data, you would go for an descriptive statistic. But if your purpose of study is to infer into the future or infer about a larger population using a smaller subset, you would go for inferential statistic. To do understand inferential statistic, you need to understand what is the concept of a population and sample.

Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 04
Introduction and Types of data Part – 2

So, the next part we are going to understand data. This is extremely important for us because statistics relies on data and when we say data it is information that is all around us, whether we are formally doing a statistical analysis or not; all of us are either creating data or contributing towards collection of data or we are collecting data ourselves. There are so many times when starting with the simple household accounts data which we keep every day.

(Refer Slide Time: 00:52)

Statistics for Data Science - I
└ Understanding data

What is Data

Date: March 7, 2020

groceries	- ₹ 200
Petrol	- ₹ 100
Snacks	- ₹ 75

In order to learn something, we need to collect data.

Definition

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation.

► Statistics relies on data, information that is around us.

We decide on how much we spend on say every day, we have people who maintain accounts and accounts could be of the kind that I come every day, I write down a date and then afterwards I say groceries 200 rupees, then petrol rupees 100.

I put the date say I put a date which is say March-7-2020 and then afterwards I will say I might have spent something on snacks; I have spent rupees 75 on snacks. Again I go back and I come back after 2 days, I do March 9-2-2020. I again calculate my data, I have again data; all this is also a data. We all collect data all the time. There is a lot of

data and we are contributing to data also every time we click the button or a keyboard or the mouse, we are generating some data.

So, as I said the definition of statistics has changed drastically over the years. So, has the nature of data. What data would have meant about 50 years back is just about numbers and categorical data. Today people talk about social media analytics, multimedia analytics, text analytics and there is so much data; even a comment on a YouTube video or a multimedia video or a photo is data. You can see that comments that come on a product e-commerce portal that is data.

So, there is data which is being collected as we speak. There is so much data that is generated there is so much data that is being collected and there is so much data waiting to be processed into meaningful information. That is the purpose so and whenever we want to do a statistical analysis, we rely heavily on data.

So, first let us define what is data, very simply put data is just facts and figures collected; by facts, it could be numerical, it could be any type. So, I just said that the comments on a multimedia or a video or anything on the internet or a product all of this contribute to what we call data.

So, what is data? Data is fact; it is what is there. We want to summarize this data we want to analyze this data for presentation and interpretation. It is very wrong to say that I want my data to tell something. You do not want the data to say anything, you in fact, the data is a fact you use data to extract information what the data is telling for interpretation purposes. So, this is the knowledge we need to understand what this data is all about and in this module, we are going to understand what is data.

(Refer Slide Time: 03:59)

Statistics for Data Science - I
└ Understanding data

Why do we collect Data

Car Model Statistical Retail Biology Mixed

- Interested in the characteristics of some group or groups of people, places, things, or events.

State	Population	Heavy
AP		
Telangana		
Orissa		
WB		
??		

(Refer Slide Time: 04:08)

S.No	Player Name	Jersey No	Matches played	Role	Runs	Batting Avg	Highest score	Wickets	Bowling Avg	Best bowling
2	Sachin Tendulkar	10	463	Batsman	18426	44.83	200	154	44.48	5/32
3	Virat Kohli	18	248	Batsman	11867	59.34	183	4	166.25	1/15
4	MS Dhoni	7	350	Batsman, WK	10773	50.58	183	1	31	1/14
5	Rohit Sharma	45	224	Batsman	9115	49.27	267	8	64.38	2/27
6	Sehwag	48	251	Batsman	8273	35.04	219	98	40.14	4/6
7	Gambhir	5	147	Batsman	5238	39.68	150	0	0	0/13
8	Yuvraj	12	304	All-rounder	8701	36.56	150	111	45.48	5/31
9	R Jadeja	8	165	All-rounder	2296	31.89	87	187	44.8	5/36
10	Zaheer Khan	34	200	Bowler	792	12	34	282	29.4	5/42
11	H Singh	3	236	Bowler	1237	13.3	49	269	33.30	5/31
12	Bumrah	93	64	Bowler	39	3.7	10	104	24.43	5/27
13	M Shami	11	77	Bowler	145	7.74	25	144	25.42	5/89
14	R Ahire	99	111	Bowler	675	16.06	65	150	39.21	4/25
15	Kuldeep Yadav	23	60	Bowler	118	13.31	19	104	26.16	6/25
16	Y Pathan	6	42	Bowler	5	2.5	3	55	24.35	6/25
17	Harik Pandya	33	54	All-rounder	951	29.91	83	54	40.64	3/31
18	Kedar Jadhav	81	73	All-rounder	1389	42.09	120	27	27.78	3/23
19	KD Karthik	21	94	Batsman, WK	1752	30.2	79	—	—	—
20	Robin Uthappa	6	46	Batsman	934	25.94	86	—	—	—
21	Ambati Rayudu	5	55	Batsman	1694	47.05	124	3	41.33	1/5
22	Rahul Dravid	19	344	Batsman	10889	39.16	153	4	42.5	2/43

Why do we collect data? Now we go back to our cricketing data set why do we collect why is this data collected at all? What are the, what is the purpose of this data? We might want to answer questions that given this data, we would like to answer questions that who is the person who has played the highest number of matches. This is a very valid question to ask or who has the highest batting average or we would want to know who is a person who has taken the highest number of wickets.

So, you see that or I would want to know here of course, I have only Sachin Tendulkar has played matches 463 matches, but then afterwards I would want to seek data about every match Tendulkar has played of this 463 matches to see that how has his batting performance been over every match and where he has played that match to see whether there has been a difference in his in house or in country batting performance to out of country batting performance. I might want to do that with every player. So, as I speak, you can see that we are generating a lot of questions and for to answer all these questions what we need is data.

So, what is the why do we collect data? We collect data the primary reason why we collect data is we are interested in knowing about the characteristics of groups; it could be groups of people, it could be places, it could be things, it could be events. Notice that we are not always interested only in people.

For example, I could have just a collection of data wherein I have a car model, number of doors of that car, whether it is diesel, whether it is petrol, whether it is an electric car. So, here you see I have absolutely no people involved or group of people involved, how many then what is the mileage of the car, whether it is a sedan or whether it is a hatchback; all these things are something which I am going to collect.

So, you see immediately whenever I want to talk I am saying groups of people it could be things, there could be a data which I am collecting wherein I have this state and the population, literacy rate. So, the state Andhra Pradesh, Telangana, Assam, West Bengal, Tamil Nadu; I could note down the population and what is the literacy rate.

Here you see I am interested in knowing about the states so, it could be places; it could be anything. So, when we are restricting ourselves it need not be only to people. So, why do we collect data? We collect data whenever we are interested in some characteristic or attribute and we seek data to answer about these characteristics or attributes.

(Refer Slide Time: 07:30)

Why do we collect Data



- ▶ Interested in the characteristics of some group or groups of people, places, things, or events.
- ▶ Example: To know about temperatures in a particular month in Chennai, India.
- ▶ Example: To know about the marks obtained by students in their Class 12.
- ▶ To know how many people like a new song/product/video-collected through comments.



Example, I would also want to know about the temperature in a particular month in Chennai. So, Chennai is a place again I am not interested about people here, I would want to know about the marks obtained by students, I might also want to know how many people like a new song. These days any internet or anything that is streaming over the internet you will find likes and dislikes. So, you might want to know how many people like a new song, new product, new video. This is collected in an entirely different way through comments.

So, as we speak we see that the data is when I talk about temperatures, you can see that there is a way I collect this data. When I talk about marks, there is a way I collect the data, marks could be either percentages or it could be grades. When I talk about marks, marks need not always and nowadays most of the boards and colleges do not give percentage marks. There are lot of people who have switched over to grades or letter grades to evaluate students.

So, you can see even that constitutes data and comments; comments is also data, but that is a completely different form of data, it is textual data. So, the minute we talk about data we see that data is all around us. We need to understand that we are collecting data to actually because we are interested in characteristics of groups or people or events. So, this is why we collect data.

(Refer Slide Time: 09:26)

Statistics for Data Science -I
└ Understanding data

Data collection



- ▶ Data available: published data.
- ▶ Data not available: need to collect, generate data.

We assume data is available and our objective is to do a statistical analysis of available data.



The next thing is, where do we collect this data from? Where do I get this data from? To answer this question, you can see that either you go and collect the data; you need to collect data and generate data or there is already data available, there is published data which is available everywhere ok. One site which you can always go and look at most of the governments publish their own data sites.

(Refer Slide Time: 09:57)



So, if you look at; so, this is a site data dot gov. It is an open government data platform India. So, this data site almost gives about all the data that is collected at a government

level. So, you can see that you have drinking water and sanitation, health data, you have economy data, the transport data, education data; so there is a data available here.

So, the key idea here what I want to convey is data is either you go and collect data or you have published data you can work on any data. If the questions you are seeking to answer needs data that has to be generated or collected; you have to go and collect the data.

This course is not going to lead you to understand how to collect this data or generate this data, but we assume that the data is available to us and our objective is to do is statistical analysis of available data. That is the purpose or the objective in this course. Nevertheless if you are seeking answers to questions for which data is not available, you need to understand how to collect this data in a structured or a scientific way how to generate this data. This will also be taught to you in due course.

(Refer Slide Time: 11:41)

Statistics for Data Science -1
└ Understanding data

INSTITUTE OF TECHNOLOGY
ROORKEE

Customer 1: Maggie, KitKat, Pepsi, Colgate toothbrush
Customer 2: Maggie, Coke, toothbrush
Customer 3: Tea, Bisleri water, biscuits

Now, when I come to data as I said that suppose I have a file; I have a file which is of this kind I just ask a person in a perhaps a retail market and I ask him what are the things that have been sold. And he comes up with a data of this kind customer 1 bought Maggie, KitKat, Pepsi and perhaps some Colgate toothbrush. Customer 2 bought Maggie, Coke, toothbrush. Customer 3 has bought say tea, then Bisleri water, they bought some cookies.

So, this is the data the person sitting at a retail counter is just collecting and if they present to you something of this kind, you cannot make any meaning out of this data which is presented to us. In a sense that this is data nevertheless this is data to you, but can we make any meaning out of this data? Suppose imagine this person who has collected this customer data is giving us similar data for 50 customers. Ok?

So, immediately this is data, but it is in an unstructured form. We have not given any or it is in an unorganized form this is data nevertheless, but it is not in a very structured or in an organized form. So, in this course we are looking to only analyze data that come to us in a structured form.

(Refer Slide Time: 13:48)

Statistics for Data Science -I
└ Understanding data

Unstructured and structured data



- ▶ For the information in a database to be useful, we must know the context of the numbers and text it holds.
- ▶ When they are scattered about with no structure, the information is of very little use.
- ▶ **Hence, we need to organize data**



So, for information of a data to be useful we must know the context of the numbers and text it holds. When they are scattered just as the example I described earlier, it is with no structure the information is of very little use, but; however, I need to organize data. So, the compelling need for me now is to organize data. This is the most important thing which I need to do.

What do I mean by organizing data?

(Refer Slide Time: 14:25)

The slide has a header 'Statistics for Data Science - I' and 'Understanding data'. It features a logo of the Institute of Technology, Anna University. The main content is a bulleted list about datasets:

- ▶ A structured collection of data.
- ▶ it is a collection of values - could be numbers, names, roll numbers.
- ▶ <https://docs.google.com/spreadsheets/d/15nJvZ-xBZDGb0oi-NcvSIY4fETotXcJdm5pV1Fq2aI/edit?usp=sharing>
- ▶ https://docs.google.com/spreadsheets/d/1qZWmXsIpFx10srpFcni9DPA961UMBtxkC1Ur_SxByq4/edit?usp=sharing
- ▶ <https://docs.google.com/spreadsheets/d/1lrmhe-E0A2LWpTB9cBK9dm-sL2SPVXYZ10MJHI6vqhM/edit?usp=sharing>

A woman in a blue sari is visible on the right side of the slide.

So, I look at a structured collection of data.

(Refer Slide Time: 14:38)

The screenshot shows a Google Sheets document with the title 'Lec0_student data'. The spreadsheet contains a table with the following data:

S.No	Name	Gender	Date of Birth	Marks in Clas Board (Board)	Marks in Cla Board (Class 12)	Mobile Number
1	Anjali	F	17 Feb, 2003	484 State Board	394 CBSE	xxx7252826
2	Pradeep	M	3 Jun, 2002	514 ICSE	437 CBSE	xxx5243748
3	Varsha	F	2 Mar, 2001	527 CBSE	442 CBSE	xxx5242824
4	Divya	F	22 Mar, 2003	397 State Board	401 State Board	xxx6546889
5	Thomas	M	19 Dec, 2002	562 CBSE	451 CBSE	xxx4242736
6	Sarita	F	19 May, 2002	533 ICSE	462 ICSE	xxx5242577
7	Prashant	M	30 Oct, 2001	496 CBSE	413 CBSE	xxx3352630
8	Harsha	M	11 Feb, 2001	436 CBSE	375 CBSE	xxx1702736
9	Rafiq	M	31 Jul, 2002	501 ICSE	423 CBSE	xxx0026248

So, if you look at a structured collection of data, we could see this is the data set which you are already in this is the data set we were already looking at when we will this is the data set we have already looked at in the last when we introduced the course. So, here you can see again this is a hypothetical data set. Imagine if this data set were collected by a person when every student was entering the college and all that the person was doing was as a person was entering the college wrote down Anjali, female, board, what is the

board ICSE, marks obtained 98 and all of that then afterwards Ramu, male, and all of this if that person had done this kind of a data collection, then again it would have been unstructured.

So, what we look at is we are trying to give a structure to this data in terms of what we refer to as a data table and in this course, we are going to largely use Google sheets to analyze our data. Hence we want to put this data or there are many ways of tabulating your data.

We are going to now describe what is the way we are going to use to analyze data in this course and mainly we are going to use spreadsheets this Google sheets to analyze a data. So, we are going to understand now how we are going to structure or organize data in a spreadsheet, how would we go about it that is the next thing which we are going to organize.

(Refer Slide Time: 16:24)

S.No	Date(Date-Month)	Time (IST)	Height (cm)	Gender	Weight (Kg)	Blood Group	Body Temperature (F)	Blood Pressure
1	02-03	7:30	178	M	75	O+	100	118/80
2	02-03	8:00	150	F	57.5	A-	98.4	125/85
3	02-03-2020	8:12	162	M	61	O-	98.2	120/80
4	02-03	8:52	145	M	65	B+	78.5	123/82
5	02-03	9:00	180	M	72	A+	95.5	109/86
6	02-03	9:09	175	M	98	O+	110	155/95
7	02-03	10:00	157	M	69	B-	94	116/80
8	02-03	10:10	165	F	59	O-	93	115/80
9	03-03	7:40	169	M	65	A+	96	130/85

(Refer Slide Time: 16:32)

S.No	A	B	C	D	E	F	G	H	I	J
12	11	03-03	8:01	156 M		61 O-		98.9	126/82	
13	12	03-03	8:15	158 F		52 B+		96.7	135/85	
14	13	03-03	8:41	183 M		82 AB-		102	123/82	
15	14	03-03	9:00	167 M		71 B-		90.9	134/89	
16	15	03-03	9:30	169 M		63 A+		94.5	118/79	
17	16	04-03	7:20	171 M		70 AB+		97.5	115/76	
18	17	04-03	8:27	163 F		67 O-		98	121/83	
19	18	04-03	9:45	155 F		64 B-		95.7	115/75	
20	19	04-03	9:56	150 M		55 A+		100	117/77	
21	20	04-03	8:39	145 F		58 AB-		94.6	122/83	

If you look at another data set here you can see that this data set has collected over some 9 patients; sorry it has collected about 20 patients who are entering a diagnostic center over a period of time 7:30, 8:00, 8:12; the height, gender, blood group, body temperature and blood pressure. So, again you see that this gives us a sort of tabulated data. Now we are going to understand what we mean by a structured collection of data and that is what we are going to focus on now.

(Refer Slide Time: 17:11)

NAME	FEES PAID	MARKS/100
Anguli	30100	75
Reena	31200	83
Lalitha	31200	92
Deepak	30100	66

When I mean by a structured collection of data to form a data set, I first need to understand, what is a variable.

Now, suppose again I have a set of data which looks like this I have name and I am looking at the fees paid, the course is just a BSc course say BSc Computer Science course, assuming. This is again hypothetical data. I have names again Anjali, Bernard, Callum, Deepak etcetera.

When you look at the fees paid again these are hypothetical numbers. If I am looking at 30000 all in INR 30000, 30000, 30000; if I am looking at fees paid and the next thing I am going to look at is what are the marks obtained out of 100 I could have 75, 83, 92, 66.

Now, when I look at this table, you can see that when I look at fees paid all of them are paying the same fees. So, there is absolutely nothing that I want to ask when it comes to fees paid. Everybody is paying the same fees, nothing is changing it is a constant along all these people. But whereas, I look at the marks I can see Anjali has obtained 75, Bernard has obtained 83, Callum has obtained 92 and Deepak has obtained 66; in other words this is varying.

I have a concept of variability there. So, when we look at what is a variable, the answer is very simple here again that we are just introducing at a basic level what is a variable. So, you can see that I can define a variable as the following.

(Refer Slide Time: 19:29)



Variables and cases

- ▶ Case (observation): A unit from which data are collected
- ▶ Variable:
 - ▶ Intuitive: A variable is that "varies".
 - ▶ Formally: A characteristic or attribute that varies across all units.
- ▶ In our school data set:
 - ▶ Case: each student
 - ▶ Variable: Name, marks obtained, Board etc.
- ▶ Rows represent cases: for each case, same attribute is recorded
- ▶ Columns represent variables: For each variables, same type of value for each case is recorded.

I can define a variable as something that “varies” and formally it is a characteristic or attribute that varies across all units.

Now, let us understand what is a unit and what is a variable in each of the data sets I have described so far. Now if we go to this data set, you can see that name is a variable in a sense, gender is definitely a variable it is not taken from a if I were taking this data from a men's-only college or a girls-only school, then this would not have been a variable; it would have been a I would have had everybody from the same gender. But here you can see that it is variable, the marks obtained is variable.

So, is the date of birth not everybody was born on the same day it could be likely again. It could be very likely if the year of birth was taken perhaps it was not varying a lot, but again it would definitely there would be in some light amount of variability there; the board, the mobile number will come and see.

So, here if you look at it the way we have defined name, gender, date of birth, marks, the board and all of them are variables whereas, Anjali, Pradeep, Varsha, Divya they are all cases or observation. On each case on each case, I have each variable recorded for each case for Anjali I have each variable recorded.

Similarly I come to the hospital data. The variable is time the date; if you look at the date you see, it was over the same day. So, here we can see that date is not varying, but here yes it is third. The first eight observations if my data is a subset and I am looking only at the first eight observations, it was taken on the same date. So, that is not varying whereas, time of entries varying. This is also data; height varies, gender varies, blood group varies, so is the blood pressure and body temperature.

So, you can see that in this case the person I have not noted down their number, but I can call them the person 1 is recorded at time of arrival, height, gender, weight, blood group, body temperature and blood pressure. Similarly each person who enters the system is recorded on each of the variables.

Similarly in the players data set, I have jersey number. Now interestingly you see jersey number also no two players have the same jersey number and you see that everybody has a different jersey number, the matches played, the role, country, here that could also be a

variable, but since this is only India then country is not a variable; highest score, wickets, bowling average.

Now, interestingly in this data set, you find some data of this kind. You can find in this there is a data which is telling 0 and there is also a data which is just showing dashes. Now what does this mean? You see that Gautam Gambhir did not take any data, but he has bowled, ok? So, the value is 0 whereas, when you look at the data of Dinesh Karthik or Robin Uthappa, you see that they do not have the data that is available for giving them any bowling statistics at all.

So, in this case even though I am collecting data, there it could be quite possible that the data which I am seeking a subset of variables which I am seeking might not be available for every unit as we have seen in this case, ok?

So, this is what we refer to this data is not available, nevertheless these people are a very much part of the data set. They do have a batting average ok, but their bowling averages are not available. What I want to emphasize at this point is this non-availability of data is different from a data taking a value 0; it means a lot. This 0 is taken even though a person has bowled 13 overs whereas, here the data is not available because these people have not bowled. We cannot take it as 0, it would mean a completely different story at this point of time.

So, when we look at data the first thing which we need to understand is what is an observation and what is a variable. So, intuitively a variable is that varies formally it is a characteristic that varies across all units. If the characteristic is available for that unit, as we saw in the cricket data set that characteristic is not available for certain players, the characteristic of bowling averages.

In our school data set, each student was a case the variable was name, marks obtained, board etcetera, rows represent cases for each case same attribute is recorded either it is recorded or we say not available. In the cricketing data set, the attribute of bowling, bowling average was not available for certain players. So, even though we record as not available if that attribute is not available, we record it as 0 if it is a value 0, there is a difference between a value 0 and not available.

Columns represent variables and for each variable the same type of value is recorded. What do I mean by same type of value? Again let us go back to our hospital data set. In this hospital data set, there are two variables here, one is which is the height variable and one is which is the weight variable.

Next to height, you see a centimeter which is written and next to weight you can see that a kilogram is written. Now suppose and again you look at it at the data collection starts at 7:30 and ends at 10 O'clock on 2nd March. If I have people who are working in shifts of 2 hours duration, a person who starts at 7 ends at 9 and a person who starts at 9 goes up to 11. The person who comes in for a shift at 9 decides to take the person's height and mention it in feet. So, he would she would or he or she would be mentioning 180 centimeters as 6 feet rather than 180.

(Refer Slide Time: 27:25)

S.No	Date(Date-Month)	Time (IST)	Height (cm)	Gender	Weight (Kg)	Blood Group	Body Temperature (F)	Blood Pressure
1	02-03	7:30	178	M	75	O+	100	118/80
2	02-03	8:00	150	F	57.5	A-	98.4	125/85
3	02-03-2020	8:12	162	M	61	O-	98.2	120/80
4	02-03	8:52	145	M	65	B+	78.5	123/82
5	02-03	9:00	6	M	72	A+	95.5	109/86
6	02-03	9:09	5'11"	M	98	O+	110	155/95
7	02-03	10:00	4'9"	M	69	B-	94	116/80
8	02-03	10:10	165	F	59	O-	93	115/80
9	03-03	7:40	169	M	65	A+	96	130/85

So suddenly your data set would start appearing after a certain point as 6. You would start looking at a data set. So, this would be 6; this might be some 5 feet 11 inches and things like that 4 feet 9 inches.

Now, you see immediately and then at 10 O'clock another person comes and the again restart. So, you see that these three data units even though they are measuring height, there is no consistency in the units that have been used. Now this as we look at a data set a primary glance of the data set itself tells us there is some problem with the data set.

Whenever we measure data, measuring a variable if it has units; we had to be consistent about the units we are using across all the observations. And that is what we mean by saying that columns represent variables and for each variable, the same type of value for each case is recorded. Again by type of value what do I mean? I cannot go back to this data set and where for example, in my data set here I have date of birth which is 17 February 2003.

(Refer Slide Time: 29:01)

S.No.	Name	Gender	Date of Birth	Marks in Cla Board (Board)	Marks in Cla Board (Class 12)	Mobile Number
1	Anjali	F	17 Feb, 2003	484 State Board	394 CBSE	xxx7252826
2	Pradeep	M	3 Jun, 2002	514 ICSE	437 ICSE	xxx5243748
3	Varsha	F	2 Mar, 2001	92% CBSE	442 CBSE	xxx5242824
4	Divya	F	22 Mar, 2003	397 State Board	401 State Board	xxx6546889
5	Thomas	M	19 years	562 CBSE	451 CBSE	xxx4242736
6	Sarita	F	19 May, 2002	533 ICSE	462 ICSE	xxx5242577
7	Prashant	M	21 years months	496 CBSE	413 CBSE	xxx3352630
8	Harsha	M	11 Feb, 2001	436 CBSE	375 CBSE	xxx1702736
9	Rafiq	M	31 Jul, 2002	501 ICSE	423 CBSE	xxx0026248

Now, suddenly I might not want to change for some people; I might not want to come and say 19 years or 21 years 3 months. Technically speaking both of them are capturing in some sense if I know the date of birth, I can compute what is age; if I know the age, I can compute what is the date of birth. But they are not the correct way of representing data.

So, when I am computing or collecting data especially in the format I wanted to I need to ensure the following that the rows represent each case and columns represent variables and for each variables, I ensure that same type of value. The another example let us go back again here, I have marks obtained in the class so, 484. I cannot suddenly say this is 92 percent or something even though that is technically it is an evaluation right.

(Refer Slide Time: 30:28)



We have organized data in a spreadsheet into a table

- Each variable must have its own column.
- Each observation must have its own row.



So, that is something which we need to take care of that every; so, when I talk about a data set what I need to be very careful and what I need to understand is I have a data set for which I know each variable has its own column, I have defined what is a variable.

If the variable has units, then every observation has its own row and every observation has the variable or each variable is measured for every observation, the units are consistent. I cannot have an observation taking the unit of height which is a variable in centimeters and another observation which is giving the unit of height in feet. I need to have the variable height taking the same type of value for each observation.

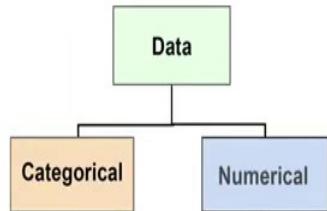
So, at this stage we understand what is data and the data set for this course is the data set which would be organized in a spreadsheet. We have shown a couple of examples here. This is the school data organized in a spreadsheet, this is the hospital data which is again organized in a spreadsheet, these are the players data which is organized in the spreadsheet. What we need to understand and remember is the columns represent the variables, the rows represent observations or cases.

For every observation I am marking what is the variable value. If the variable value is not available, I put a not available symbol; I capture the non availability of that particular variable for that observation. So, at this point of time, we have a data set available for us for analysis. So, the next step we need to understand is what do we understand about this variable, how do I classify these observations; that is the next thing.

(Refer Slide Time: 32:34)

Statistics for Data Science -1
└ Classification of data
 └ Categorical and numerical

Categorical and numerical



Navigation icons: back, forward, search, etc.

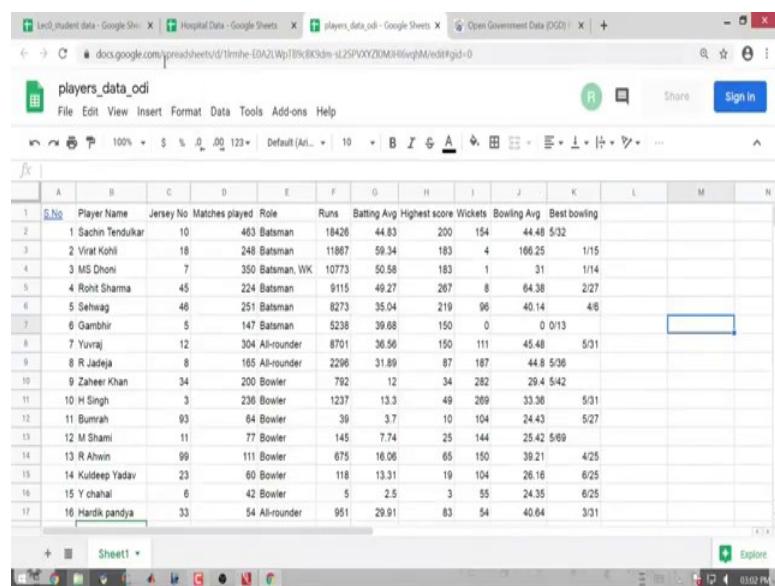


Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 05
Introduction and Types of data Part – 3

If you are come up to this module what would I expect you to know is you know what is data and you know how the data is organized as a data table.

(Refer Slide Time: 00:26)



S.No	Player Name	Jersey No	Matches played	Role	Runs	Batting Avg	Highest score	Wickets	Bowling Avg	Best bowling
1	Sachin Tendulkar	10	463	Batsman	18426	44.83	200	154	44.48	5/32
2	Virat Kohli	18	248	Batsman	11867	59.34	183	4	166.25	1/15
3	MS Dhoni	7	350	Batsman, WK	10773	50.58	183	1	31	1/14
4	Rohit Sharma	45	224	Batsman	9115	49.27	267	8	64.38	2/27
5	Sehwag	46	251	Batsman	8273	35.04	219	98	40.14	4/6
6	Gambhir	5	147	Batsman	5238	39.68	150	0	0.0/13	
7	Yuvraj	12	304	All-rounder	8701	36.56	150	111	45.48	5/31
8	R Jadeja	8	165	All-rounder	2298	31.89	87	187	44.8	5/36
9	Zaheer Khan	34	200	Bowler	792	12	34	282	29.4	5/42
10	H Singh	3	236	Bowler	1237	13.3	49	269	33.38	5/31
11	Umrah	93	64	Bowler	39	3.7	10	104	24.43	5/27
12	M Shami	11	77	Bowler	145	7.74	25	144	25.42	5/69
13	R Ahwin	99	111	Bowler	675	16.06	65	150	39.21	4/25
14	Kuldeep Yadav	23	60	Bowler	118	13.31	19	104	26.16	6/25
15	Y chahal	6	42	Bowler	5	2.5	3	55	24.35	6/25
16	Hardik pandya	33	54	All-rounder	951	29.91	83	54	40.64	3/31

And when I say it is organized as a data table I mean that you understand that this is the data where I have columns representing variables and rows representing the cases or observations.

(Refer Slide Time: 00:39)



Now, once we go back to this, you can see that again let us go back to the data set. The minute you look at a data set of this kind, you can see that when I look at name it is just Anjali, Pradeep, Varsha, Divya, I have gender. When I look at gender I have two categories female and male. When I have marks, you can see that this marks. Let me put back this mark here, let me put it as say 565 and remove the percentage.

So, when I look at marks, you can see that there is 484, 514, 565 etcetera, but state board again it is again some sort of a category here. I have a State Board, I have ICSE, I have CBSE. Marks in class 12 again 394, 437 again the board etcetera.

(Refer Slide Time: 01:35)

S.No	Date(Date-Month)	Time (IST)	Height (cm)	Gender	Weight (Kg)	Blood Group	Body Temperature (F)	Blood Pressure
1	02-03	7:30	178	M	75	O+	100	118/80
2	02-03	8:00	150	F	57.5	A-	98.4	125/85
3	02-03-2020	8:12	162	M	61	O-	98.2	120/80
4	02-03	8:52	145	M	65	B+	78.5	123/82
5	02-03	9:00	153	M	72	A+	95.5	109/86
6	02-03	9:09	167	M	98	O+	110	155/95
7	02-03	10:00	175	M	69	B-	94	116/80
8	02-03	10:10	165	F	59	O-	93	115/80
9	03-03	7:40	169	M	65	A+	96	130/85

Let us go to the other hospital data. Height, again I have centimeter. So, let me rephrase this into centimeters I put it as 152, 167 and I have a 175 centimeter here, gender again male female, weight is again you can see 75, 57.5, 65, 98. There is something called blood group, body temperature in degree Fahrenheit and you have blood pressure.

(Refer Slide Time: 02:12)

S.No	Player Name	Jersey No	Matches played	Role	Runs	Batting Avg	Highest score	Wickets	Bowling Avg	Best bowling
1	Sachin Tendulkar	10	463	Batsman	18426	44.83	200	154	44.48	5/32
2	Virat Kohli	18	248	Batsman	11887	59.34	183	4	166.25	1/15
3	MS Dhoni	7	350	Batsman, WK	10773	50.58	183	1	31	1/14
4	Rohit Sharma	45	224	Batsman	9115	49.27	267	8	64.38	2/27
5	Sehwag	46	251	Batsman	8273	35.04	219	98	40.14	4/6
6	Gambhir	5	147	Batsman	5238	39.68	150	0	0	0/13
7	Yuvraj	12	304	All-rounder	8701	36.56	150	111	45.48	5/31
8	R Jadeja	8	165	All-rounder	2298	31.89	87	187	44.8	5/58
9	Zaheer Khan	34	200	Bowler	792	12	34	282	29.4	5/42
10	H Singh	3	236	Bowler	1237	13.3	49	269	33.36	5/31
11	Bumrah	93	64	Bowler	39	3.7	10	104	24.43	5/27
12	M Shami	11	77	Bowler	145	7.74	25	144	25.42	5/69
13	R Ahrin	99	111	Bowler	675	16.06	65	150	39.21	4/25
14	Kuldeep Yadav	23	60	Bowler	118	13.31	19	104	26.16	6/25
15	Y chahal	6	42	Bowler	5	2.5	3	55	24.35	6/25
16	Hardik pandya	33	54	All-rounder	951	29.91	83	54	40.64	3/31

So, the minute you look at the data of this kind. Let us look at the cricket data, you again have a bowling average which is 44.48, 166.25, 64.38 whereas, matches played you have 463, 248, 350. Now, here you have a jersey number which is 10, 18, 7 etcetera. So, the

minute the you look at a data set of this kind two things you observe. The first thing you observe is you have numbers, you have text. By text I mean that you have names, you have role of a batsman, all rounder, go back to your blood group you have O plus A minus, you have gender which is again captured as male and female, you have board again which is captured as CBSE state board.

So, you immediately see that all data is not of the same kind. You see that there is a basic difference in the way the data is presented. So, the next thing is we seek whether I can classify this data into broadly two categories.

Immediately I notice that some data is numerical in nature or quantitative in nature. For example, I see that the marks and the bowling averages the height, weight, the bowling average, the highest score, the batting average, the matches played, etcetera can be clubbed into some kind of variable whereas the name the gender, the board, the blood group and here the role. These represent certain kind of variables.

Interestingly if you look at this jersey number, it appears to be numbers; but all of us know that these numbers have no meaning. So, there are kinds of variables which could take numerical values as in this thing they could be numbers, but they might not.

So, there is definitely a difference between the variable, a jersey number and matches played even though both are numbers. So, it is very important for us to understand how to classify data and to what category or what type of data my variable belongs to extremely important. So, when we look at data, data is broadly classified into two categories; categorical data and numerical data.

(Refer Slide Time: 05:17)

Statistics for Data Science -1
Classification of data
Categorical and numerical

Categorical and numerical variables



- ▶ Categorical data
 - ▶ Also called qualitative variables.
 - ▶ Identify group membership
- ▶ Numerical data
 - ▶ Also called quantitative variables.
 - ▶ Describe numerical properties of cases
 - ▶ Have measurement units
- ▶ Measurement units: Scale that defines the meaning of numerical data, such as weights measured in kilograms, prices in rupees, heights in centimeters, etc.
 - ▶ The data that make up a numerical variable in a data table must share a common unit.

So, when we look at categorical data, these are also called as qualitative variables. Now, it identifies group membership. What do we mean by group membership? Again we go back to our student data, let us look at gender. Gender is a categorical variable. I have two categories here. I can classify any observation into one of these two categories.

So, it is a group membership. Similarly, when I look at board I have a category which is a State Board, I have ICSE, I have CBSE. So, again you can see that this categorical variable has three categories and any observation can be categorized into one of these three groups.

So, when we go back you can see that you in a sense, I am giving membership of an observation to a particular group in that particular variable. So, this category has groups. Let us go to the hospital data. You see that blood group every patient is either an O positive or an O negative or a B positive or a A positive or A negative.

So, you can see that there are many blood groups I again this is a categorical variable; gender is a categorical variable. What kind of variable is mobile number? I leave it as an exercise. I want you people to come up with an answer; mobile numbers are again numbers. What kind of a variable do you think is a mobile number?

Similarly, what kind of a variable do you think is the jersey number? Even though jersey number is 10, 18, 7, what kind of a variable is it? Is it a categorical variable or is it a numerical variable?

So, now, the first thing which we need to understand is I have categorical data I also have what are numerical data. When we have numerical data, numerical data is also called quantitative variables. Here I can talk about numerical properties of data.

Now, go back here. Marks obtained both in class 10 and class 12 are numerical data and you talk about marks this is 484 marks, this is 514 marks, this is 565 marks. Come to the hospital data I have 178 centimeters, 150 centimeters; weight 75 kilograms, 575 sorry 57.5 kilograms; body temperature in terms of Fahrenheit degree Fahrenheit 100 degree Fahrenheit, 98.4 degree Fahrenheit. When I come to cricketing data matches played, I have batting averages, I have wickets taken 154 run score 200 runs.

So, you can see immediately when I talk about numerical data, I have associated with them either measurement units or I have something which are called the bowling average and batting average. Now, you also see that when I talk about matches, it is a whole number it is 463 whereas, when I talk about batting averages you can see that it can take any value. It could be fractions also or it is any value. So, this again tells us that when we talk about categorical and numerical data; within numerical data, I could have discrete data and I could have continuous data. I could further look at data that is discrete and I could look at data that is continuous right.

So, once we understand what is this categorical and numerical data, we need to understand the measurement units that are used for numerical data. Again let us go back to our data, you can see that here height is measured in centimeter, weights is measured in kilograms, the body temperature in degree Fahrenheit. Again we have marks again this is 484 marks. Again when you come to players data you have matches played, highest score; it is in runs wickets taken again in wickets ok.

So, the idea is we need to understand what is the scale that defines the numerical data. Again we have already emphasized on the point that when you have numerical data which take units. We need to ensure that the variable is measured across all observations and shares a common unit. This is something which we need to ensure.

(Refer Slide Time: 11:15)

Statistics for Data Science -1
└ Classification of data
 └ Cross-sectional versus time-series data

Cross-sectional and time-series data



- ▶ Time series - data recorded over time
- ▶ Timeplot – graph of a time series showing values in chronological order
- ▶ Cross-sectional - data observed at the same time

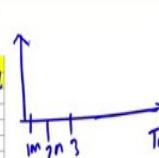


Apart from categorical data and numerical data, we also have data which where which are referred to as time series data.

(Refer Slide Time: 11:27)

Statistics for Data Science -1
└ Classification of data
 └ Cross-sectional versus time-series data

Time-series data- Example



Date	Qty(kg)	Potato cost (Rs.)	Selling price(Rs.)
01-Mar	0	21	24
02-Mar	1300	20.05	24
03-Mar	675	20.5	24
04-Mar	0	NA	NA
05-Mar	675	20.8	24
06-Mar	675	21.25	24
08-Mar	20	20.5	24
09-Mar	900	20.5	24
10-Mar	900	20.5	24
11-Mar	0	NA	NA
12-Mar	900	20.3	24
13-Mar	1125	19.4	22
15-Mar	1125	18.8	22
16-Mar	1125	19.4	22
17-Mar	1125	19.25	22
18-Mar	1125	20.3	24
19-Mar	1125	19.8	24
20-Mar	675	21.25	24
22-Mar	675	20.5	24
23-Mar	0	NA	NA
24-Mar	0	NA	NA
25-Mar	675	19.6	24
26-Mar	675	19.7	24
27-Mar	1125	19.3	24
29-Mar	540	20.6	26
30-Mar	0	28	



Let us look at this data. Now, when we look at this data, you can see that this data has a variable which is called date. And this data is actually the data which tracks at a retail outlet, what is the quantity that is procured every day from 1st March to 30th March of a month, the cost of procurement and the price at which it was being sold.

If you look at it the variable is just one thing which is potato and you for all the days you are tracking what is the quantity that is being sold. So, in other words I have a date; I have the first march, I have second march, I have third march. I have a time and over that time, I can actually find out what is the quantity that is being procured every day.

So, this is what we refer to as a time series data where the data on a particular variable; this could be the quantity procured on potato is obtained the variable is the same. What is the variable? It is quantity of potatoes that is the variable, but I am tracking this variable over a period of time which is from 1st March to 30th March. So, this kind of data is what we refer to as a time series data whereas, cross sectional data is the data which is observed at the same time.

(Refer Slide Time: 13:24)



The image shows a navigation bar with the following structure:

- Statistics for Data Science -1
- Classification of data
- Cross-sectional versus time-series data

A horizontal blue line extends from the right side of the "Cross-sectional versus time-series data" link to the right edge of the slide. To the right of this line is the logo of the Institute of Technology, featuring a circular emblem with a lamp and the text "INSTITUTE OF TECHNOLOGY". Below the navigation bar is the word "Summary".

- ▶ Classify data as categorical or numerical.
- ▶ For numerical data, find out unit of measurement.
- ▶ Check whether data is collected at a point of time (cross-sectional data) or over time (time-series data).



So, we will broadly classify, we should know that given data we classify them broadly as categorical or numerical. So, whenever we are presented with a data set, we should be able to classify all the variables in the data set as a categorical variable or a numerical variable. If it is a numerical variable, find out what is the unit of measurement. Again check if the unit of measurement is consistent across all the observations. The third thing is check whether it is collected at a point of time whether it is cross sectional data or whether it is time series data. So, now, given a data set you should be able to do this.

Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 06
Introduction and Types of data Part – 4

(Refer Slide Time: 00:16)



- ▶ Data collection requires one of the following scales of measurement: nominal, ordinal, interval, or ratio.



So, when you have come here I assume at this point of time you know what is data, you know you have a good data set. By meaning you have a tidy data set what I mean is you have actually a data set which is organized into variables and you have cases. Each case is recorded for each variable and if I do not have a data, I will not record that data. So, this is a tidy data set.

Further, from this data set you are able to look at the variables, you are able to broadly classify these variables as categorical variables and you are able to classify them as numerical variables. So, at this point of time this is what you should be able to do with your data set.

Now, what do we want to do next? Now, we again examine each of our variables in greater detail because if you look at the definition of the statistics which we gave earlier, it said that you want to learn from data. Now, when I want to learn from data the immediate thing is I want to see whether I can summarize this data. Again, when I say I want to summarize this data, the question I ask is can I come up with some graphical

summaries and can I come up with numerical summaries. The minute I say numerical summaries, I need to know whether I can do arithmetic operations on the data, ok.

So, to know whether I can do arithmetic operations on data, I need to understand what are the scales of measurement I use for my data. Now, when we look at scales of measurement, I have 4 scales of measurement and they are called the nominal, ordinal, interval and ratio scale. We are going to understand about each of these scales of measurement in great detail.

Why is it important? It is extremely important for us to know what is the scale of measurement for each of the variables I have in my data set to eventually come up with what is the kind of summary I can do for that variable. Hence, it is extremely important for us to know what is the scale of measurement for each variable.

(Refer Slide Time: 03:01)

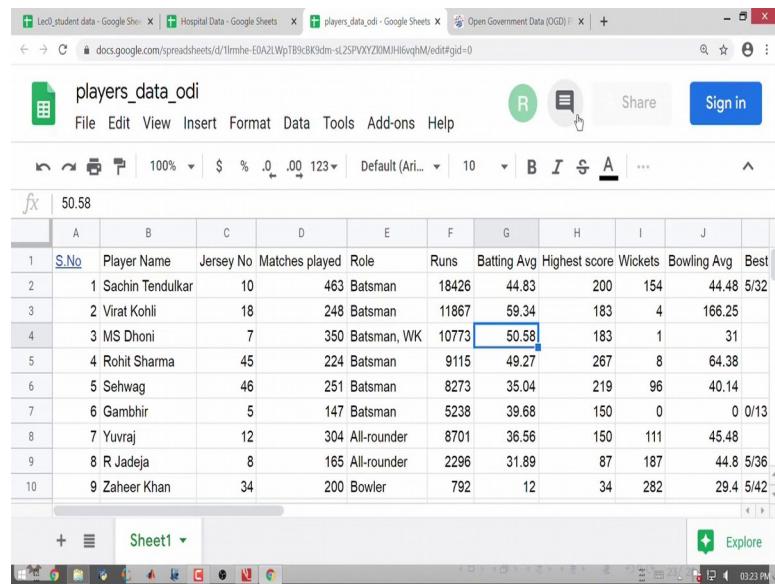


- ▶ When the data for a variable consist of labels or names used to identify the characteristic of an observation, the scale of measurement is considered a **nominal** scale.
- ▶ Examples: Name, Board, Gender, Blood group etc.
- ▶ Sometimes nominal variables might be numerically coded.



We start with the nominal scale. When the data consists of labels or names, the scale of measurement is considered a nominal scale.

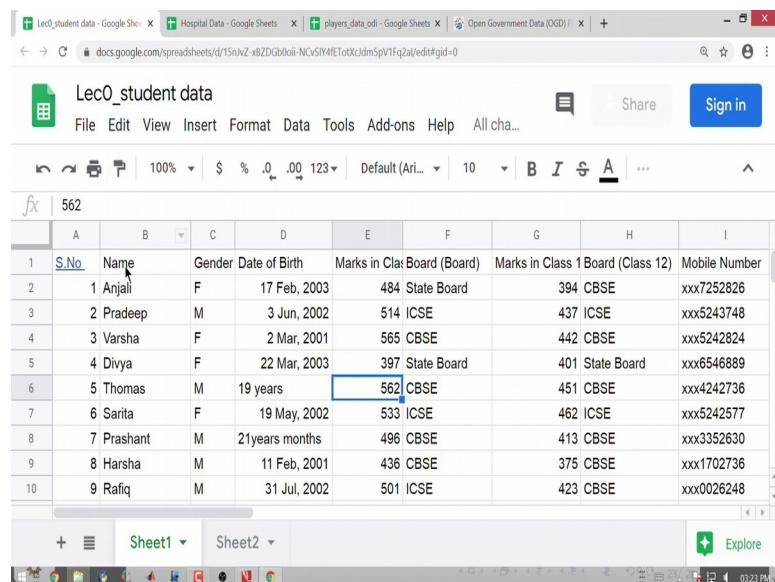
(Refer Slide Time: 03:16)



The screenshot shows a Google Sheets document titled "players_data_odi". The table has columns for S.No, Player Name, Jersey No, Matches played, Role, Runs, Batting Avg, Highest score, Wickets, Bowling Avg, and Best. The data includes rows for Sachin Tendulkar, Virat Kohli, MS Dhoni, Rohit Sharma, Sehwag, Gambhir, Yuvraj, R Jadeja, and Zaheer Khan.

S.No	Player Name	Jersey No	Matches played	Role	Runs	Batting Avg	Highest score	Wickets	Bowling Avg	Best
1	Sachin Tendulkar	10	463	Batsman	18426	44.83	200	154	44.48	5/32
2	Virat Kohli	18	248	Batsman	11867	59.34	183	4	166.25	
3	MS Dhoni	7	350	Batsman, WK	10773	50.58	183	1	31	
4	Rohit Sharma	45	224	Batsman	9115	49.27	267	8	64.38	
5	Sehwag	46	251	Batsman	8273	35.04	219	96	40.14	
6	Gambhir	5	147	Batsman	5238	39.68	150	0	0	0/13
7	Yuvraj	12	304	All-rounder	8701	36.56	150	111	45.48	
8	R Jadeja	8	165	All-rounder	2296	31.89	87	187	44.8	5/36
9	Zaheer Khan	34	200	Bowler	792	12	34	282	29.4	5/42

(Refer Slide Time: 03:19)

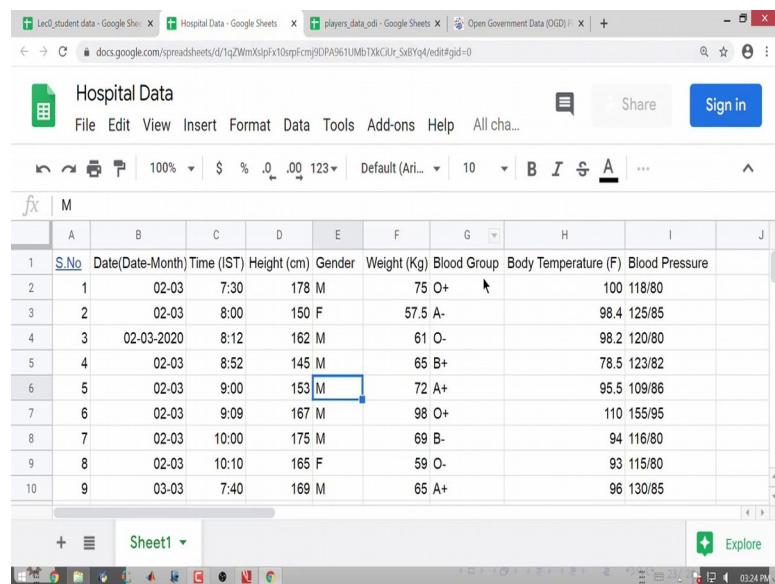


The screenshot shows a Google Sheets document titled "LecO_student data". The table has columns for S.No, Name, Gender, Date of Birth, Marks in Class 1 Board (Board), Marks in Class 12 Board (Class 12), and Mobile Number. The data includes rows for Anjali, Pradeep, Varsha, Divya, Thomas, Sarita, Prashant, Harsha, and Rafiq.

S.No	Name	Gender	Date of Birth	Marks in Class 1 Board (Board)	Marks in Class 12 Board (Class 12)	Mobile Number
1	Anjali	F	17 Feb, 2003	484 State Board	394 CBSE	xxx7252826
2	Pradeep	M	3 Jun, 2002	514 ICSE	437 ICSE	xxx5243748
3	Varsha	F	2 Mar, 2001	565 CBSE	442 CBSE	xxx5242824
4	Divya	F	22 Mar, 2003	397 State Board	401 State Board	xxx6546889
5	Thomas	M	19 years	562 CBSE	451 CBSE	xxx4242736
6	Sarita	F	19 May, 2002	533 ICSE	462 ICSE	xxx5242577
7	Prashant	M	21 years months	496 CBSE	413 CBSE	xxx3352630
8	Harsha	M	11 Feb, 2001	436 CBSE	375 CBSE	xxx1702736
9	Rafiq	M	31 Jul, 2002	501 ICSE	423 CBSE	xxx0026248

Again, let us go back to our example. You can see in the first example I have name which is evidently a nominal scale because it is only names, ok. Again board you can see that it is labels in some sense I have a state board I have a ICSE, I have a CBSE, ok. Whereas gender, again it is a label, I am labelling this category with a female and a male.

(Refer Slide Time: 03:57)



The screenshot shows a Google Sheets document titled "Hospital Data". The spreadsheet contains a table with 10 rows of data. The columns are labeled: S.No, Date(Date-Month), Time (IST), Height (cm), Gender, Weight (Kg), Blood Group, Body Temperature (F), and Blood Pressure. The data includes various patient details such as age, gender, height, weight, blood group, temperature, and blood pressure. Row 6 is currently selected, highlighting the "Gender" column.

S.No	Date(Date-Month)	Time (IST)	Height (cm)	Gender	Weight (Kg)	Blood Group	Body Temperature (F)	Blood Pressure
1	02-03	7:30	178	M	75	O+	100	118/80
2	02-03	8:00	150	F	57.5	A-	98.4	125/85
3	02-03-2020	8:12	162	M	61	O-	98.2	120/80
4	02-03	8:52	145	M	65	B+	78.5	123/82
5	02-03	9:00	153	M	72	A+	95.5	109/86
6	02-03	9:09	167	M	98	O+	110	155/95
7	02-03	10:00	175	M	69	B-	94	116/80
8	02-03	10:10	165	F	59	O-	93	115/80
9	03-03	7:40	169	M	65	A+	96	130/85

So, you can see that names, board, gender. We go back here, we have the blood data. In the blood group data you see again I can label it as O positive, A minus, O negative. One thing we need to notice here is I just have labels, I have only names; there is no particular order of these names.

For example, the data would not have made any difference that is whether I am having a female male or a male female. It is absolutely it does I am just having this as a label to identify the characteristic. So, this is the called a nominal scale of measurement. Sometimes we can see that nominal variables may be numerically coded. What do I mean by a numerically coded nominal variable?

For example, we have gender. Gender takes two labels which is male and female. I might code a male a 0 and a female a 1, I might code a male a 1 and a female is 0. So, this numerically coded is again equivalent to just labelling this variable, this numeric; no sanctity about having a 0 or a 1.

I could label a man or 2 and woman a 3, or a woman a 5 and man a 7. All it says is this label has the same understanding; these numbers have no meaning when you are coding the nominal variables. Both the codes are valid that is what we mean.

(Refer Slide Time: 05:40)

Nominal scale of measurement



- ▶ When the data for a variable consist of labels or names used to identify the characteristic of an observation, the scale of measurement is considered a **nominal** scale.
- ▶ Examples: Name, Board, Gender, Blood group etc.
- ▶ Sometimes nominal variables might be numerically coded.
 - ▶ For example: We might code Men as 1 and Women as 2. Or Code Men as 3 and Women as 1. Both codes are valid.
- ▶ There is no ordering in the variable.
- ▶ **Nominal: name categories without implying order**



When no nominal variables are coded whether 1 or 2 or 3 and 1, it has both codes are valid. There is no ordering. This is extremely important that we understand when I talk about nominal variables; there is no ordering in the variable. So, a nominal variable is just name categories without implying order.

Going back to a data sets in the student data set, name, gender, board are nominal variables. In a blood bank data set, we have gender and blood group which are nominal variables. In the cricketing data set jersey number is a nominal variable. The role of batsman is a nominal variable, that is and the player name is a nominal variable. So, nominal scale of measurement is used when I have name categories without implying any order.

(Refer Slide Time: 06:55)

Statistics for Data Science -1	
└ Classification of data	
└ Scales of measurement	
Ordinal scale of measurement	CUSTOMER 1
	1 EXCELLENT
2	GOOD
3	BAD
4	GOOD



- ▶ Data exhibits properties of nominal data and the order or rank of data is meaningful, the scale of measurement is considered a **ordinal** scale.
- ▶ Each customer who visits a restaurant provides a service rating of excellent, good, or poor.
 - ▶ The data obtained are the labels—excellent, good, or poor—the data have the properties of nominal data.
 - ▶ In addition, the data can be ranked, or ordered, with respect to the service quality.

BAD GOOD EXCELLENCE



The next scale of measurement is called an ordinal scale of measurement where the data exhibits the same property of nominal data, but here an order or rank is meaningful. What do I mean by this is for example, a customer who visits a restaurant provides a service rating of excellent good and poor. The data are again labels.

The data is again labels. For example, I can have a data where I have customer 1 who gives a rating of excellent, customer 2 good, customer 3 bad, customer 4 again good. So, if you look at the variable, the variable is rating.

Here you can again see this is taking a nominal value. By nominal it is taking a categorical value where my categorical variable has 3 categories; excellent, good and bad, but within this categorical variable there is an order. You know the order is bad, good and excellent. So, categorical or nominal data which exhibit some rank or an order or rank is meaningful is said to have the measurement, the scale of measurement is said to be a ordinal scale.

(Refer Slide Time: 08:42)

Statistics for Data Science -1
└ Classification of data
└ Scales of measurement

Ordinal scale of measurement

The diagram illustrates an ordinal scale of measurement. It shows three categories: "BAD", "GOOD", and "EXCELLENT" arranged horizontally. Below them, the numbers "1", "2", and "3" represent their respective ranks. A bracket above the categories connects "BAD" and "EXCELLENT" with the word "GOOD" in the middle. Another bracket below the numbers connects "1" and "3" with the number "2" in the middle.

- ▶ Data exhibits properties of nominal data and the order or rank of data is meaningful, the scale of measurement is considered a **ordinal** scale.
- ▶ Each customer who visits a restaurant provides a service rating of excellent, good, or poor.
 - ▶ The data obtained are the labels—excellent, good, or poor—the data have the properties of nominal data.
 - ▶ In addition, the data can be ranked, or ordered, with respect to the service quality.
- ▶ **Ordinal – name categories that can be ordered**



A photograph of a woman with glasses and a blue sari, identified as the speaker, positioned next to the slide content.

So, ordinal scale of data is name categories that can be ordered.

(Refer Slide Time: 08:53)

Statistics for Data Science -1
└ Classification of data
└ Scales of measurement

Interval scale of measurement

The diagram illustrates an interval scale of measurement. It shows three numerical values: "1", "2", and "3". A bracket is placed below these values, grouping them together with the text "numerical values that can be added/subtracted (no absolute zero)" written underneath.

- ▶ If the data have all the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure, then the scale of measurement is **interval** scale.
- ▶ Interval data are always numeric. [Can find out difference between any two values.]
- ▶ [Ratios of values have no meaning] here because the value of zero is arbitrary.
- ▶ **Interval:**
numerical values that can be added/subtracted (no absolute zero)



A photograph of a woman with glasses and a blue sari, identified as the speaker, positioned next to the slide content.

The next scale is called an interval scale of measurement. Now, when we talk about ordinal scale, again I can code an ordinal scale of measurement. For example, in our earlier example my bad could have been coded as 1, my good can be coded as 2, and my excellent can be coded as 3. I could code them.

There is an order in 1, 2, 3. But then one thing which I need to understand here is the distance between bad to good need not be the same as the distance between good and

excellent. It is just an order, I know excellent is better than good, but I cannot say that excellent the difference between good and excellent is the same as the distance between good and bad. I have an order, but at this point of time I am not able to comment anything more about this order.

So, when I go to interval scale of data, interval scale of data has all the properties of interval scale of data, but the interval between the values is expresses a fixed unit of measure. Remember, when I said bad, good and excellent, I said the difference between good and bad need not be the same as the difference between excellent and good.

Whereas, when I have internal data I have an ordering, but in this case whenever I am ordering my data the interval between the values is expressed in units of a fixed unit of measure. When the differences expressed in a fixed unit of measure; so, if I have for example, it can I can find out the difference between any two values, ok.

Now, here ratios do not have any meaning because the value of 0 is arbitrary. Let us explain this through an example. Interval data and numerical values that can be added or subtracted, it has no absolute 0.

(Refer Slide Time: 11:05)

Statistics for Data Science -1
└ Classification of data
└ Scales of measurement

Example: temperature

- ▶ Suppose the response to a question on how hot the day is comfortable and uncomfortable, then the temperature as a variable is nominal.
- ▶ Suppose the answer to measuring the temperature of a liquid is cold, warm, hot - the variable is ordinal.
- ▶ Example: Consider a AC room where temperature is set at 20°C and the temperature outside the room is 40°C. It is correct to say that the difference in temperature is 20°C, but it is incorrect to say that the outdoors is twice as hot as indoors.

Let us look at temperature as an example. Suppose, the reference or response to a question is how hot the day is, and you respond as just comfortable uncomfortable or just good, bad. I am just giving a label to my feeling. I really I am not grading or ordering

this feeling I have, whether I am just telling it is comfortable it is categorized into comfortable and uncomfortable or it could be just leisurely or the something. But in a sense you might feel that there is an order, but at this point of time I am not having any order between comfortable and uncomfortable. Temperature is a variable of interest and here you can see temperature is a nominal variable.

Now, suppose temperature is again the variable of interest, but here I am interested in knowing how hot a liquid is, whether it is cold, warm, or hot; you see that there is a order in this variable. I know cold has warm is warmer than cold or hot is warmer than warm, the variable is ordinal. However, here I do not know whether the difference between a warm and a cold beverage is the same as a hot and a warm beverage.

But now suppose I am measuring the temperature, consider an AC room which is set at 20 degree centigrade and temperature out of their outside the room is 40 degree centigrade it is correct to say that the difference in the temperature is 20 degree centigrade, absolutely fine.

Suppose, I had set it at 10 degree, 14 degree centigrade and the temperature outside was 28 degrees, it is perfectly right for me to tell that there is a difference in the temperature of 14 degree centigrade. But it is incorrect; it is absolutely incorrect for me to say that outdoors is twice as hot as indoors because 40 degree centigrade is not twice as hot as 20 degree centigrade.

So, when I am talking about an interval scale I know in an interval scale, I know the difference between 10 degree centigrade and 20 degree centigrade is 10, which is same as a difference between 20 degree centigrade and 30 degree centigrade, which is same as a difference between 30 degree centigrade and 40 degree centigrade.

But I cannot make a statement that 40 degree centigrade is twice as hot as 20 degree centigrade. It is incorrect. So, when I am able to; so, that tells me that I can talk about the difference between any two values, but here ratios have no meaning.

(Refer Slide Time: 14:13)



Example: temperature

- ▶ Suppose the response to a question on how hot the day is comfortable and uncomfortable, then the temperature as a variable is nominal.
- ▶ Suppose the answer to measuring the temperature of a liquid is cold, warm, hot - the variable is ordinal.
- ▶ Example: Consider a AC room where temperature is set at 20°C and the temperature outside the room is 40°C. It is correct to say that the difference in temperature is 20°C, but it is incorrect to say that the outdoors is twice as hot as indoors.
- ▶ Temperature in degrees Fahrenheit or degrees centigrade is an interval variable. No absolute zero.

	Celsius	Fahrenheit
Freezing point	0	32
Boiling point	100	212



Again, we understand from temperature, at least when we talk about Celsius and Fahrenheit scales there, there is no absolute 0, in the Celsius 0 and 100 are set to be as the freezing point and the boiling point whereas, in Fahrenheit it is 32 and 212. Only in the Kelvin you have a 0 degree, where 0 means absolutely no temperature. But when you are talking about Celsius and Fahrenheit we understand that there is no absolute 0.

So, when you talk about an interval scale, it is extremely important for us to understand there is no absolute 0. However, the difference between an interval scale and an ordinal scale of measurement is in an interval scale the difference between the values is fixed unit of measure whereas, for a ordinal scale that need not be a fixed unit of measure that is good to bad need not be the same difference as excellent to good. This is the key difference.

The last scale of measurement is what we refer to as the ratio scale of measurement.

(Refer Slide Time: 15:25)

Ratio scale of measurement



- ▶ If the data have all the properties of interval data and the ratio of two values is meaningful, then the scale of measurement is **ratio scale**.
- ▶ Example: height, weight, age, marks, etc.
- ▶ **Ratio: numerical values that can be added, subtracted, multiplied or divided (makes ratio comparisons possible)**



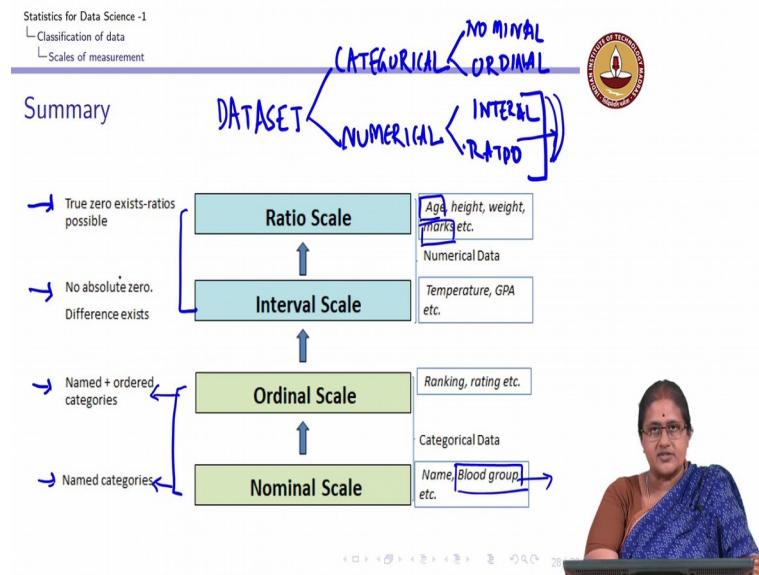
In a ratio scale of measurement it has all properties of interval data and the ratio is a very meaningful measure. The scale is a ratio scale. So, the example, height, weight, marks. Like I know a person who has scored 300 has scored twice as well as a person who has scored 150 marks.

I know a person with a score of 200 is highest score of 200 is twice as good as a person who has scored 100 runs. I know a person who has taken 100 wickets is twice as good as a person who has taken 50 wickets.

So, I can have a notion of a ratio which I can define here, the variables, height, weight, marks, runs, wickets all of them are examples of ratio scale of measurement. So, ratio when you have a variable which is measured in the ratio scale; you can do all the mathematical or arithmetic operations on it.

You can add, you can subtract, you can multiply or divide. Whereas, when you talk about interval scale you can only talk about difference or you can add and subtract. You have no absolute 0, so ratios do not have any meaning.

(Refer Slide Time: 17:04)



Whenever you are presented with the data set after you identify the variables as categorical or numerical, it is extremely important for us to understand that when I have categorical data, I have the nominal and ordinal scale. Within the nominal scale, I have nominal which is only named category, ordinal is a name with an order. Here the difference between order is not a fixed measure.

Again, example for categorical data name, blood group or nominal scale; ordinal scale ranking, rating; there is a order, but then there need not be a fixed order in the rating. Absolute 0 does not exist for an interval scale. This is for numerical data or quantitative data. Absolute 0 does not exist, but proportional difference exists.

I can have the mathematical operation of difference done here. And in the ratio scale, ratios are possible. All variables of age, height, weight, marks etcetera are variables which can be measured on a ratio scale. Temperature, grade point, average everything is a 4 GPA is not twice as good as a 2 GPA, they can be measured on a in interval scale

So, whenever we are given data why are we interested about the scales of measurement? Here you can see no arithmetic operations possible. Here I can have some sense of an order. Here I can do addition, subtraction. Here I can do all arithmetic operations. So, the minute I identify the variable the type of questions, I asked makes sense.

For example, when I have a variable which is a blood group I would not be asking the question of what is the average blood group because I cannot define any arithmetic operation here. Similarly I need to, but when I am talking about a age or a mark, I can ask about some numerical summaries depending on what is this scale of measurement.

So, with this we stop this module. At the end of this module you should be able to look at a data set which is very well organized, identify variables as categorical or numerical. And once you are able to identify these variables further look at what are the scales of measurement whether it is a nominal, whether it is an ordinal scale of measurement or whether it is an interval scale of measurement or ratio scale of measurement.

Most of the textbooks and books written in statistics clubbed; both of these scales together and mentioned it either as an interval or ratio scale. But nevertheless there is a difference, and the critical difference is in an interval there is no absolute 0, whereas in a ratio scale an absolute or a true 0 exist.

Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 07
Tutorial – 1

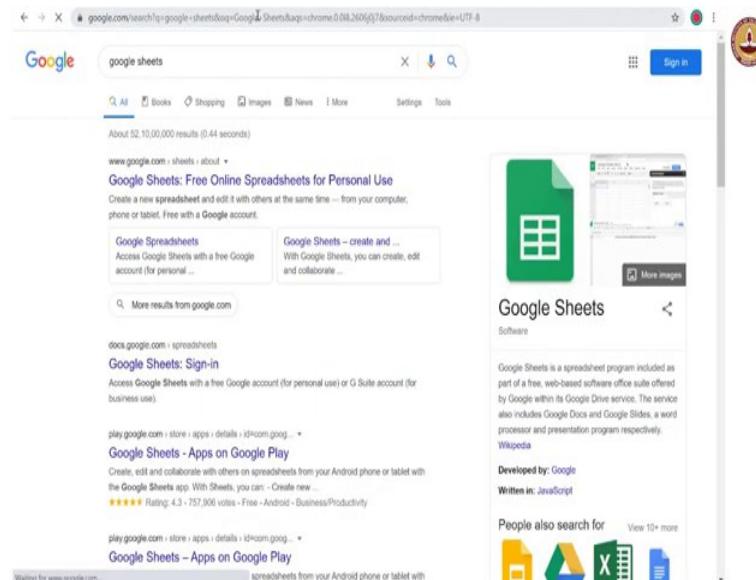
(Refer Slide Time: 00:14)

Learning Objectives for statistics week tutorials

Tutorial 1 Spreadsheet intro Learning Objectives

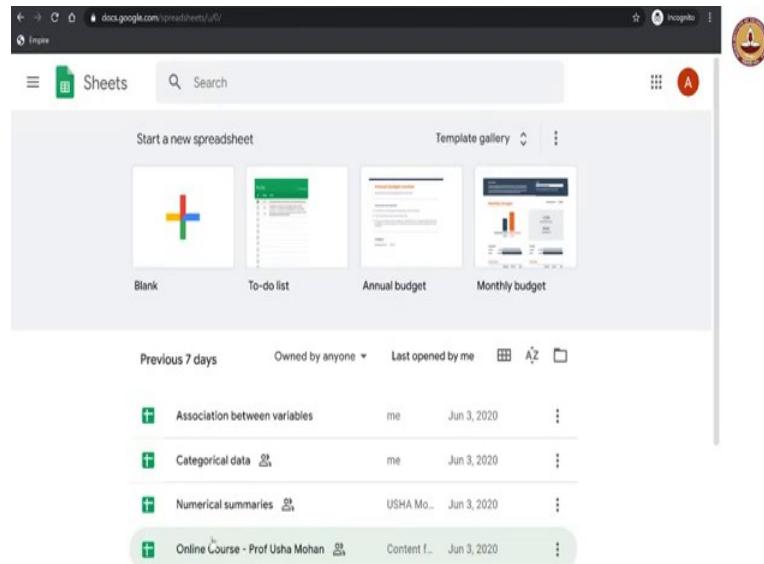
1. Creating a new blank spreadsheet in Google sheets
2. Understand the basic terminologies and notations in google sheet-like cell, column, row
3. Navigation and manipulation of cells.
4. Using a spreadsheet for calculating simple interest
5. Understanding how Autofill works in Google sheets using similar patterns from previous cells

(Refer Slide Time: 00:39)



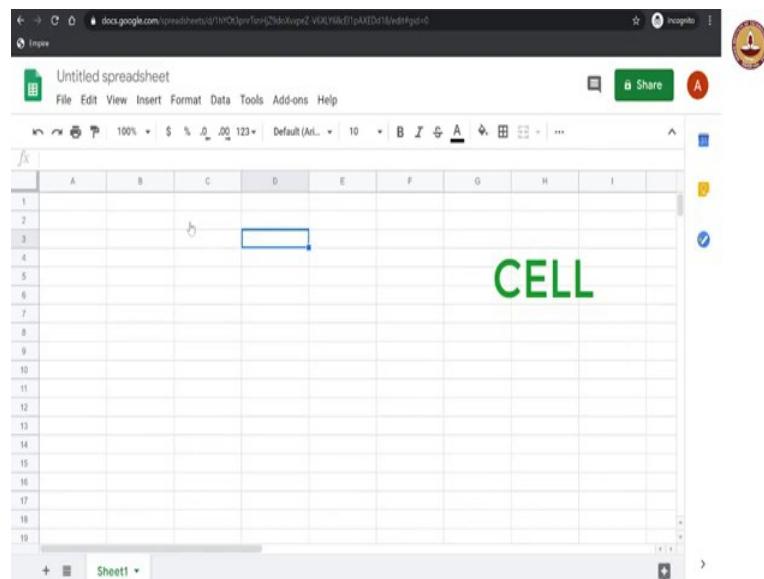
Hello, Statistics students. In this video, we are going to look at Google sheets, the very basic elements of spreadsheet. You could use a Microsoft Excel as well, but for the purposes of our course we will be doing Google sheets.

(Refer Slide Time: 00:47)



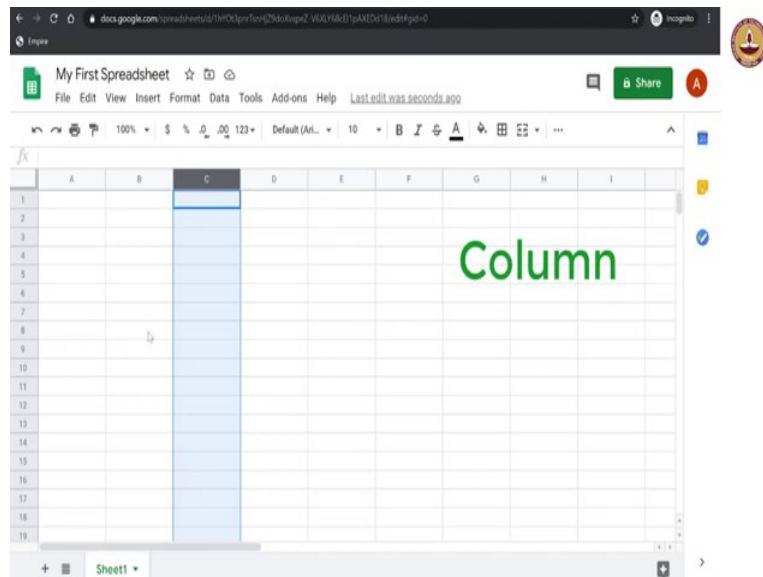
So, let us search for Google sheets. We will go to this, yeah and personal go to Google sheets. Since, I am already logged in, this is the page I get to. And now we want to start a new spreadsheet, we use this and here is our spreadsheet.

(Refer Slide Time: 01:00)



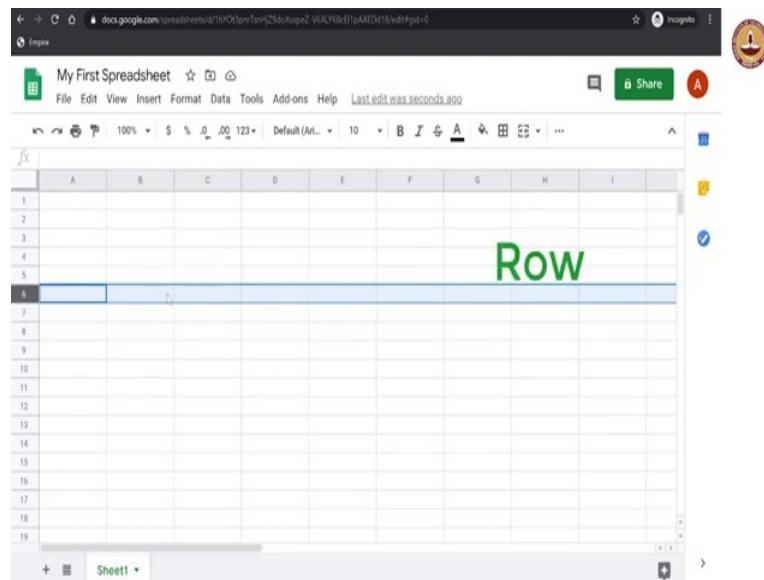
As you can see this spreadsheet has cells, each of these boxes is what is called a cell. And you could navigate through them using the arrow keys, you can go down, down, down or right or left or up. The 4 arrow keys will help you navigate across the spreadsheet.

(Refer Slide Time: 01:29)

A screenshot of a Google Sheets document titled "My First Spreadsheet". The spreadsheet has a single column labeled A through I. Column C is currently selected, highlighted with a blue background. The word "Column" is written in green text in the middle of the selected column. The top navigation bar shows standard Google Sheets menu options like File, Edit, View, Insert, Format, Data, Tools, Add-ons, Help, and a "Last edit was seconds ago" timestamp. On the far right of the top bar, there are "Share" and "A" buttons. The bottom of the screen shows the sheet tab labeled "Sheet1" and other navigation controls.

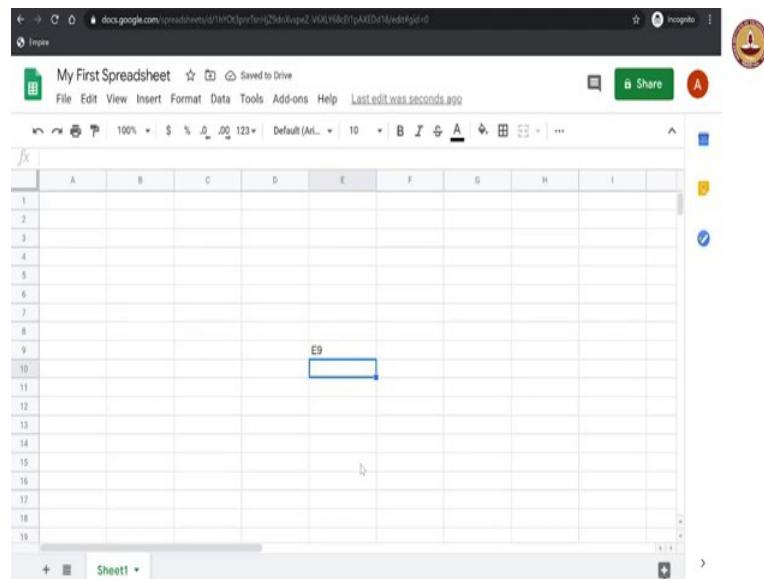
Let us give a name to our spreadsheet and we can name it here. We can go and click here. Let me call this my first spreadsheet. So, we have seen what the cells are, they are these boxes. And now these cells can be seen as a group when they are all in the same vertical region. So, this whole set of cells, the vertical set of cells is called a column. And likewise, all the horizontal set of cells here it is called a row.

(Refer Slide Time: 02:09)

A screenshot of a Google Sheets spreadsheet titled "My First Spreadsheet". The spreadsheet is currently empty, with only the header row visible. A large green text "Row" is centered in the middle of the sheet. The columns are labeled A through I, and the rows are numbered 1 through 19. The status bar at the bottom shows "Sheet1".

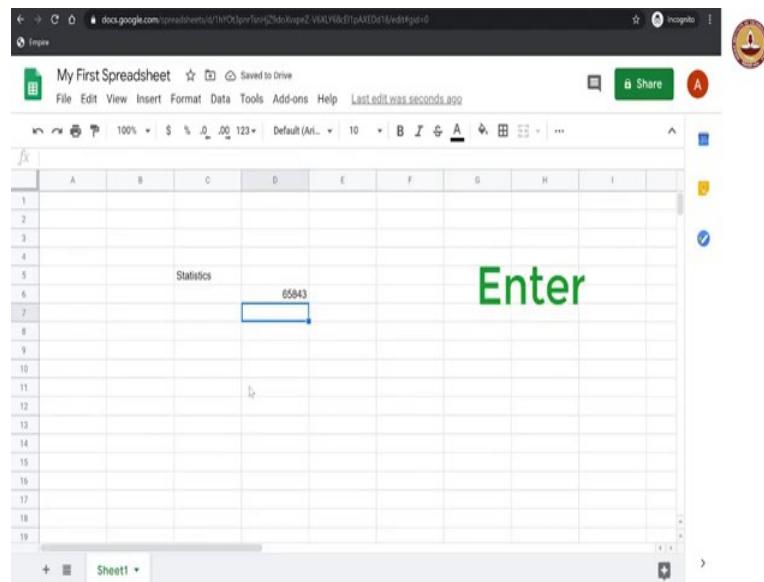
So, when I choose a particular cell let us say this one, this belongs to the E column and 9 row. So, this cell can be called E 9.

(Refer Slide Time: 02:15)

A screenshot of a Google Sheets spreadsheet titled "My First Spreadsheet". The cell E9 is highlighted with a blue border. The rest of the spreadsheet is empty, with only the header row visible. The columns are labeled A through I, and the rows are numbered 1 through 19. The status bar at the bottom shows "Sheet1".

And as you have seen, now I will go I am going to erase this. And as you have seen, you can enter data into these cells. You could enter text.

(Refer Slide Time: 02:38)

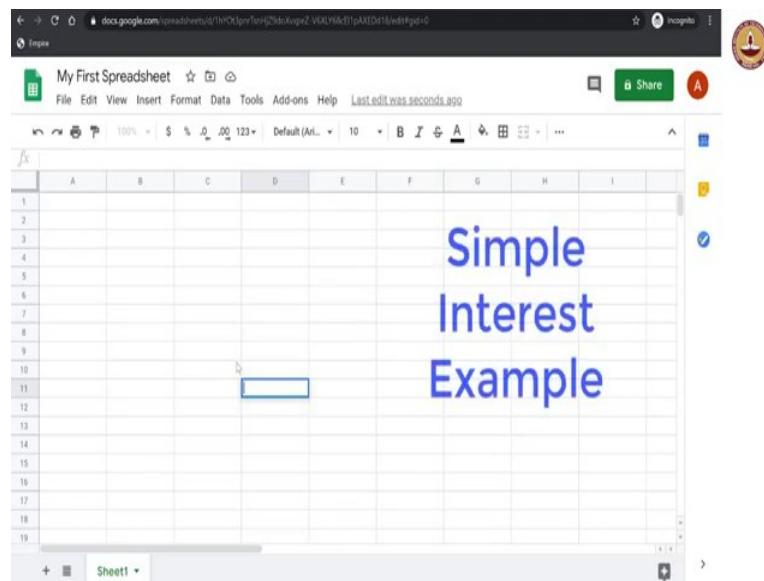


A screenshot of a Google Sheets spreadsheet titled "My First Spreadsheet". The spreadsheet has one sheet named "Sheet1". In cell A6, the word "Statistics" is written in black. In cell D6, the number "65843" is written in black. To the right of the spreadsheet, the word "Enter" is displayed in large green letters.

So, you could enter something like statistics and we press enter and it is there or you could also enter numbers. You could go back to the same cell or you can do it in a different cell and let say I enter a number 65843 enter and now the cell has that particular piece of data.

How are spreadsheets useful to us? So, for that let us take a tiny example. Let us first delete these elements. Yeah, spreadsheet is empty now.

(Refer Slide Time: 03:15)



A screenshot of a Google Sheets spreadsheet titled "My First Spreadsheet". The spreadsheet has one sheet named "Sheet1". In cell A11, the words "Simple Interest Example" are written in blue. The cell containing the text is highlighted with a blue border.

Let us say we want to make a spreadsheet for calculating simple interest over a loan.

(Refer Slide Time: 03:21)

My First Spreadsheet

Principal
= 10000

Let us say you have given a 10000 rupee loan to somebody and there is a monthly interest of 0.5 percent.

(Refer Slide Time: 03:26)

My First Spreadsheet

Rate of
Interest = 0.5%
per month

And let say you given it in the month of March.

(Refer Slide Time: 03:31)

A	B	C	D	E	F	G	H	I
1 March	0	0		10000				
2 April	50							
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								

And what is the interest for the month of March? Nothing; 0 right, because you have just given it. And so, how much is the total interest? Still 0. And then, how much is the total that you are owed? That would be 10000 rupees.

And now this is going to change when we look at the month of April. In April, the person who is taken loan from you will be required to pay an interest of 0.5 percent of 10000 rupees. And how much is that? That is 50 rupees. And that means, the total interest is 50 rupees and the total money you are owed is 10,050, right.

(Refer Slide Time: 04:22)

A	B	C	D	E	F	G	H	I
1 March	0	0	10000					
2 April	50	50	10050					
3 May	50	100	10100					
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								

And now let us look at the month of May. Now, again in this month what is the interest that is owed? That is 50 rupees again. And the total interest has now become 50 plus 50, 100. Thus the total money that is owed to you is 10100, ok. I think I mistyped it. It has to be this yes, right.

So, this way you could store the data that you have. And for each month, you can see what is the money owe or how much is the interest that is expected or how much is the total interest that is to be paid to you.

However, you do not need to really input all the data, all the cells by yourselves. The spreadsheet software gives you this very amazing utility called autofill. So, let me tell you what that is.

(Refer Slide Time: 05:50)

The screenshot shows a Google Sheets spreadsheet titled "My First Spreadsheet". The data is as follows:

	A	B	C	D	E	F	G	H	I
1	March	0	0	10000					
2	April	50	50	10050					
3	May	50	100	10100					
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									

A green text overlay on the right side of the sheet reads "Click, hold, and drag." A cursor is shown over the cell containing the value "50" in row 3, column B, indicating the start of the drag operation.

We know that for each month the interest is supposed to be 50 rupees, yes. So, I can take this particular cell and you look at this option here, right, the cursor becomes a cross. I am taking it to the corner, now I click it; hold it and I pull it down. And you see that the spreadsheet software, Google sheets itself directly fills all these cells with the same value, right. So, you do not have to type it.

And what is more interesting is the autofill can also catch patterns. So, you see this pattern, right, this is 0, this is 50, this is 100. So, there is an increment of 50. And let us

select all of these. How do you select multiple cells together? You select them using the shift key.

(Refer Slide Time: 06:33)

A screenshot of a Google Sheets spreadsheet titled "My First Spreadsheet". The spreadsheet has three columns: Month (A), Value (B), and Total (D). The data starts at row 1 with March, 0, 10000. Rows 2 and 3 show April and May respectively, with values 50 and 100 in column B. A fourth row is partially visible. The cell B2 is currently selected. A large watermark-like graphic in the center-right of the sheet area shows the text "Shift + ↓" with arrows pointing down and right, indicating the keyboard shortcut for selecting multiple cells. The bottom status bar shows "Count: 9".

So, you go to that top cell, you hold shift down arrow, down arrow. And now, you again grab that corner and you drag it down. And you see that the autofill is filling these cells with the increment of 50 each. So, you could do the same thing with this column as well, shift down, down and then hold and drag, right. Very convenient, is not it.

(Refer Slide Time: 07:02)

A screenshot of the same Google Sheets spreadsheet after using the Shift key. The cell B3 is now selected. The data now includes rows for June through January, with values increasing by 50 each month. The bottom status bar shows "Count: 9".

Month	Value	Total
March	0	10000
April	50	10050
May	50	10100
June	50	10150
July	50	10200
August	50	10250
September	50	10300
October	50	10350
November	50	10400
December	50	10450
January	50	10500

And it is quite smart. You do not even need to hold all 3 of these months, you can just take May and then hold and drag and there you go. Your spreadsheet software fills these cells for you because these are all simple patterns that it recognizes, ok.

With that we have just seen the utility of spreadsheets. We will see in more detail in the next video.

Thank you.

Statistics for Data Science - 1
Prof. Usha Mohan

Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 08

Tutorial – 2

(Refer Slide Time: 00:16)

Learning Objectives for statistics week tutorials

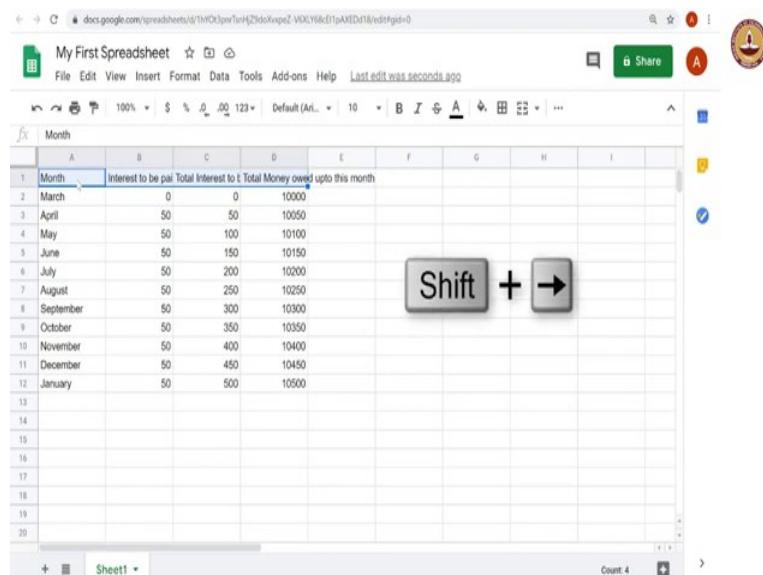
Tutorial 2 Spreadsheet Formatting Learning Objectives

Doing the following operations to format a spreadsheet so that it is

understandable to a third person

1. Labelling the columns
 2. Highlighting the label row(or headings) using a different colour for that row and formatting text by choosing bold.
 3. Formatting the text in rows and columns by using text wrap, alignment of text in the vertical and horizontal direction.
 4. Formatting the numbers into the currency format or other formats
 5. Understanding the other formatting options like italic, underline and strikethrough

(Refer Slide Time: 00:17)



We will continue with our simple interest example. But before we go into the finer utility of spreadsheet software's, you should first try to represent our data in a more meaningful

format. Right now we just put down the numbers and the month names, but if somebody were to look at this spreadsheet, they are not really going to understand what is happening right.

So, we are going to try to format this properly and for that I think, the first thing we should be doing is to label the columns that we have here. Clearly, the first column is all month and we decided the second column is the interest for that month, the third column C is the total interest and the column D is the total money that you are owed.

So, in order to represent this, it will be nice to have a heading; a label for the column. So, something above this, for that we will add a row above the first row; so, that we can keep the headings or the labels on them.

(Refer Slide Time: 01:38)

A screenshot of a Google Sheets spreadsheet titled "My First Spreadsheet". The spreadsheet contains two rows of data:

C	D
0	10000
50	10050
100	10100
150	10150
200	10200
250	10250
300	10300
350	10350
400	10400
450	10450
500	10500

The text "Right Click" is overlaid on the spreadsheet. A context menu is open over the first row (row 1), specifically over cell D1. The menu includes options like Cut, Copy, Paste, Paste special, Insert 1 above, Insert 1 below, Delete row, Clear row, Hide row, Resize row, Group row, Ungroup row, Get link to this range, Define named range, and Protect range.

And for this, we first go to the row 1, we right click and here you see a lot of options you could do. You could delete the row, you could clear the row, you could hide it, resize it, group and a lot of things.

(Refer Slide Time: 01:51)

C	D	E	F	G	H	I
0	10000					
50	10050					
100	10100					
150	10150					
200	10200					
250	10250					
300	10300					
350	10350					
400	10400					
450	10450					
500	10500					

What we are interested in is inserting 1 row above and there we go we have a row here and let us give the names. This one is month; this one is interest to be paid for the month. As you can see, it is spilling over we will attend to that in a bit and then go to this cell and this cell would be the heading for column C which is again I press enter.

This would be the total interest to be paid up to this month, again its spilling over, does not matter, we will come back to that and then here again go and press enter and this one would be the total money owed up to this month.

So, now we have a decent indication of what that column is. However, we have trouble because this does not look good. First of all, the month looks the heading looks pretty much the same as the rest of the cells in that column and here this is cut off, this is cut off and this one is spilling over. So, we are going to have to format this so, it looks better; so, it looks meaningful.

What we could do is, first we should make the headings the labels look distinct. So, we could probably try to fill them with a different color instead of the whole column, I am just going to take the non-empty ones. So, I have selected the cells A1, B1, C1 and D1 by using the shift and right key.

(Refer Slide Time: 03:56)

A screenshot of a Google Sheets spreadsheet titled "My First Spreadsheet". The sheet contains a table with columns labeled A through I. Row 1 is a header row with labels "Month", "Interest to be paid", "Total Interest to Date", and "Total Money owed upto this month". Rows 2 through 12 contain data for each month from March to January, showing increasing values for interest and total owed. The word "Fill Colour" is overlaid in green text in the middle-right area of the sheet.

Month	Interest to be paid	Total Interest to Date	Total Money owed upto this month
March	0	0	10000
April	50	50	10050
May	50	100	10100
June	50	150	10150
July	50	200	10200
August	50	250	10250
September	50	300	10300
October	50	350	10350
November	50	400	10400
December	50	450	10450
January	50	500	10500

And now I am going to go to this, this is fill color; fill color as and it will fill the cells with that particular color and I can choose whatever color. There are lot of colors here as you can see; you can choose whichever and this is a Google provided theme. If you use just these colors, you will get something that looks reasonably good. So, you can just choose that; however, I am going to go for something kind of light. Let us look at light green 3 and there you go. Now we have these cells have a distinct color.

(Refer Slide Time: 04:34)

A screenshot of a Google Sheets spreadsheet titled "My First Spreadsheet". The sheet contains the same table as the previous screenshot. The word "Bold" is overlaid in green text in the middle-right area of the sheet. Below it, the keyboard shortcut "Ctrl + B" is shown in large, stylized letters.

Month	Interest to be paid	Total Interest to Date	Total Money owed upto this month
March	0	0	10000
April	50	50	10050
May	50	100	10100
June	50	150	10150
July	50	200	10200
August	50	250	10250
September	50	300	10300
October	50	350	10350
November	50	400	10400
December	50	450	10450
January	50	500	10500

We can do better; we can make them B, bold text. So, this option is the bold text which you could also get through control B; there it is bold.

(Refer Slide Time: 05:06)

Month	=Interest to be paid	Total Interest	Total Money owed upto this month
March	0	0	10000
April	50	50	10050
May	50	100	10100
June	50	150	10150
July	50	200	10200
August	50	250	10250
September	50	300	10300
October	50	350	10350
November	50	400	10400
December	50	450	10450
January	50	500	10500

(Refer Slide Time: 05:12)

Month	=Interest to be paid	Total Interest	Total Money owed upto this month
March	0	0	10000
April	50	50	10050
May	50	100	10100
June	50	150	10150
July	50	200	10200
August	50	250	10250
September	50	300	10300
October	50	350	10350
November	50	400	10400
December	50	450	10450
January	50	500	10500

And now, we still have the trouble that some of this text is cut off some and this text is spilling over. So, what do you want to do in these cases is again select those and you can do what is called text wrapping. So, that is an additional option which is there.

So, if these three dots indicate more options and go to this and here, this is text wrapping and there are three options; one is overflow which is what you are seeing now; it is

spilling over into the next cell and this is wrapping which means that text will be adjusted into the same cell in multiple lines and this one is clip, it will get clipped to that particular end of the cell. I think the wrap is best because we want to see the whole text and there we go.

Now, you can see the whole text, it is coming up as multiple lines right and instead of three lines, we might want it to be maybe just two lines. So, for that it helps if the cell width is greater right; if the cell were wider, then we can fit this into two lines.

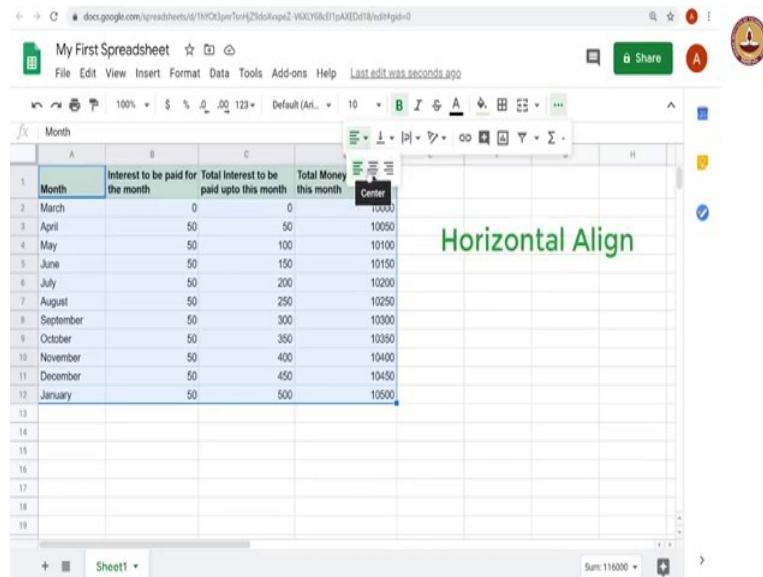
(Refer Slide Time: 06:18)

Month	Interest to be paid for the month	Total Interest to be paid upto this month	Total Money owed upto this month
2 March	0	0	10000
3 April	50	50	10050
4 May	50	100	10100
5 June	50	150	10150
6 July	50	200	10200
7 August	50	250	10250
8 September	50	300	10300
9 October	50	350	10350
10 November	50	400	10400
11 December	50	450	10450
12 January	50	500	10500
13			
14			
15			
16			
17			
18			
... n			

So, one way to do that is to expand the whole column because if you are expanding this cell, you are expanding every cell under it. So, what you can do is take your cursor to this edge, this border of the column click and drag it and as you can see, it is now expanded. You can do the same for column C and again for column D.

Now, since everything is fitting into two lines, you have the cell height to be accommodating only two lines and you could make this furthermore centrally aligned. So, what we can do is, let us select everything we can take this shift right, right, right and then, holding shift still go down, down, down, down in this way.

(Refer Slide Time: 07:17)



Month	Interest to be paid for the month	Total Interest to be paid upto this month	Total Money this month
March	0	0	10000
April	50	50	10050
May	50	100	10100
June	50	150	10150
July	50	200	10200
August	50	250	10250
September	50	300	10300
October	50	350	10350
November	50	400	10400
December	50	450	10450
January	50	500	10500
13			
14			
15			
16			
17			
18			
19			

Now, we have selected all these cells and what we can do now is, we can center align the text so, we again went to the more options and the horizontal alignment, we are doing it as center.

So, what you have seen here is that, all the data is now fitted at the center of the cell and further, we can go to vertical align as well. Here again you have a top, middle and bottom so, we go to middle which has made the text appear at the vertical center. Now, our spreadsheet looks considerably better to read.

We can do a little more formatting in terms of these numbers. These numbers could just be anything right 10,000, 10,050 etcetera. But we know that, they are supposed to be money, these are Indian rupees. So, we can do number formatting, we will again select these cells. You see what I am doing right shift right, right and then down, down, down yeah.

(Refer Slide Time: 08:107)

A screenshot of a Google Sheets document titled "My First Spreadsheet". The table has columns labeled "Month" and "Interest to be paid this month". The cell for March contains "0". The "Format" menu is open, and the "Number" section is selected. A submenu is open under "Number", showing options like "Bold", "Italic", "Underline", and "Strikethrough". The "Font size" option is also visible. The "Align" option is expanded, showing "Merge cells", "Text wrapping", "Text rotation", "Conditional formatting", and "Alternating colors". The "Merge cells" option is highlighted. The "Text rotation" option is also visible.

Month	Interest to be paid this month
March	0
April	50
May	50
June	50
July	50
August	50
September	50
October	50
November	50
December	50
January	50

(Refer Slide Time: 08:31)

A screenshot of a Google Sheets document titled "My First Spreadsheet". The table has columns labeled "Month" and "Interest to be paid this month". The cell for March contains "0". The "Format" menu is open, and the "Text rotation" option is selected. A submenu is open under "Text rotation", showing "None", "Tilt up", "Tilt down", "Stack vertically", "Rotate up", "Rotate down", and "0° angle". The "0° angle" option is highlighted. The "Text wrapping" option is also visible.

Month	Interest to be paid this month
March	0
April	50
May	50
June	50
July	50
August	50
September	50
October	50
November	50
December	50
January	50

(Refer Slide Time: 08:33)

The screenshot shows a Google Sheets spreadsheet titled "My First Spreadsheet". The table has two columns: "Month" and "Interest to be paid this month". The "Interest to be paid this month" column contains values from 0 to 50 for each month from March to January. A context menu is open over the cell for March, showing the "Format" menu. The "Number" option is selected, and a submenu is displayed with various number formats like "Plain text", "Number", "Percent", "Scientific", etc.

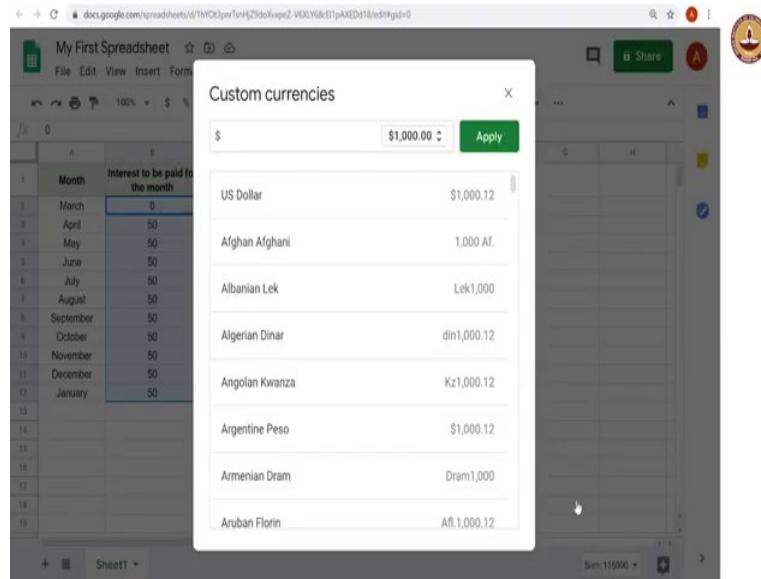
And now, we go to the format option here and here, we have a number of formatting options whatever you have seen there are also here. There is a alignment, left center right is for horizontal, top middle bottom is for vertical and there are lot of other options. We will see them slowly later.

(Refer Time: 08:36)

The screenshot shows a Google Sheets spreadsheet titled "My First Spreadsheet". The table has two columns: "Month" and "Interest to be paid this month". The "Interest to be paid this month" column contains values from 0 to 50 for each month from March to January. A context menu is open over the cell for March, showing the "Format" menu. The "Number" option is selected, and a submenu is displayed with various number formats like "Plain text", "Number", "Percent", "Scientific", etc.

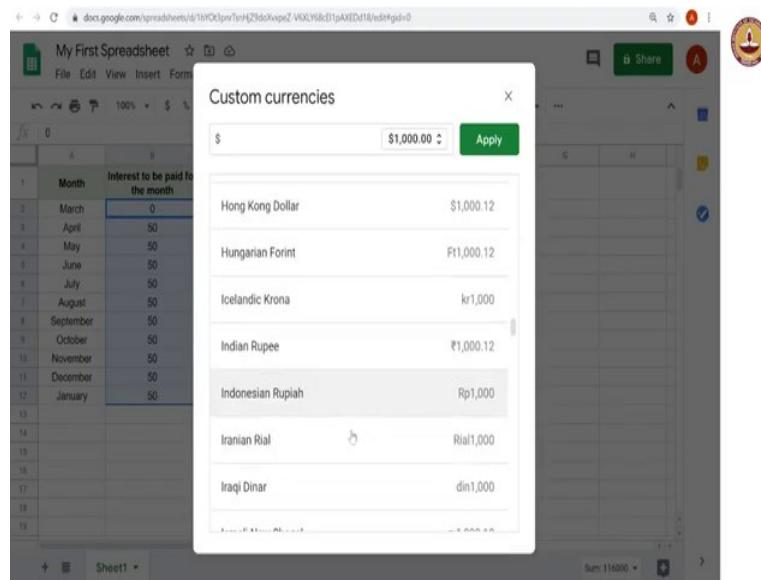
But what we are looking at is number formatting. So, in number formatting, there are lot of options. What we are looking for is currency, but the default currency that Google sheets offers is the dollar symbol; what we want is the Indian rupee.

(Refer Slide Time: 08:53)



So, we go to more formats, then more currencies and within this, if we went down we will see so, there are all these other currency options for formatting.

(Refer Slide Time: 09:04)



But, let us look for the Indian rupee and here we are. This is the rupee symbol and if you apply this, all of this is now in Indian rupees. And now this looks like a good simple interest sheet.

In this video, we have seen some of the formatting options, but they are actually a lot of them. For example, we have not looked at the Italic option or the strikethrough option

and there is also an underlining the text option which is here in the format. So, you are encouraged to go through all these options, play around with all the formatting and find formats that work for you that you think looks good. But more than looking good, make sure that your spreadsheet is doing what you wanted to do; it should represent the data to you in a very usable way.

So, the other options could be the whole we have seen text wrapping, vertical alignment, horizontal alignment. There is also the borders, fill color we have seen, this is the text color. You can play with all of this bar and the other things that are in here. We will see more of these maybe in future. Right now let us go on to seeing how to make more use of spreadsheets with the same example.

Thank you.

Statistics for Data Science - 1
Prof. Usha Mohan

Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 09

Tutorial - 3

(Refer Slide Time: 00:14)

Learning Objectives for statistics week tutorials

Tutorial 3 Spreadsheet Formulae Learning Objectives

1. Understand the basic mathematical operations that can be done in a spreadsheet.
 2. Using cell reference to refer to a particular cell to do mathematical calculations
 3. Understanding how autofill happens with respect to changing the cell numbers
 4. Fixing the column or row number in the formula while giving cell reference
 5. Fixing the cell value in the autofill so autofill uses the same cell value for other cells
 6. Using Paste Special to just copy the format of cell
 7. Changing the basic inputs of the data to change the entire calculations by using cell reference

(Refer Slide Time: 00:17)

We have formatted the sheet in the previous video, now I have removed that data. We will look at better ways to fill in those cells. Earlier, we did it by our own calculation. But this time around, let us make the spreadsheet software do the calculations for us.

So, the initial entry here we will have to do in the month of March, there is no interest the total interest will be 0 and the principal amount which is the 10,000 rupees is what is owed up to that month and as you can see, after the formatting the rupee symbol gets added by itself.

(Refer Slide Time: 01:08)

Month	Interest to be paid for the month	Total Interest to be paid upto this month	Total Money owed upto this month
March	₹0.00	₹0.00	₹10,000.00
April			
May			
June			
July			
August			
September			
October			
November			
December			
January			
13			
14			
15			
16			
17			
18			
19			

Now, for the next one, we should input the an interest for the month of April. We know that the principal is 10000 rupees and the rate of interest per month is 0.5 percent.

(Refer Slide Time: 01:26)

My First Spreadsheet

File Edit View Insert Format Data Tools Add-ons Help Last edit was seconds ago

100% Default (A1)

A B C D E F G H I

	Month	Interest to be paid for the month	Total Interest to be paid upto this month	Total Money owed upto this month					
1	March	₹0.00	₹0.00	₹10,000.00					
2	April	=							
3	May								
4	June								
5	July								
6	August								
7	September								
8	October								
9	November								
10	December								
11	January								
12									
13									
14									
15									
16									
17									
18									
19									

'=' symbol to indicate a formula

'=' symbol
to indicate
a formula

If we did the calculation, we will get it to be 50 rupees, but let us make the spreadsheet software do the calculation. And for that, for making the spreadsheet software evaluate things, we go to the cell and we press “is equal to” symbol. This tells the spreadsheet software that whatever is there in the cell, the formula that you enter needs to be calculated.

So, this would be equal to the principal amount which is $10000 * 0.5\%$, 0.5% is 0.5 divided by 100. You can see that, whatever you are entering in the cell is also in this bar which is where the cell function is being evaluated and it is already showing us that the quantity will be 50 rupees. So, once I have entered this, I press enter and we have 50 rupees.

(Refer Slide Time: 02:20)

Month	Interest to be paid for the month	Total Interest to be paid upto this month	Total Money owed upto this month
March	₹0.00	₹0.00	₹10,000.00
April	=10000*0.5/100		
May			
June			
July			
August			
September			
October			
November			
December			
January			
13			
14			
15			
16			
17			
18			
19			

(Refer Slide Time: 02:25)

Month	Interest to be paid for the month	Total Interest to be paid upto this month	Total Money owed upto this month
March	₹0.00	₹0.00	₹10,000.00
April	=10000*0.5/100		
May			
June			
July			
August			
September			
October			
November			
December			
January			
13			
14			
15			
16			
17			
18			
19			

Also pay attention to the multiplication symbol being the asterisk. The asterisk is the symbol above the 8 key. You can press shift and 8 and you will get the asterisk symbol; I am going to delete the extra one here. So, now we have done this, enter and we have 50.

As before, we are going to use auto fill to fill these cells and now they are all 50 rupees as we know the interest for each month is a fixed amount and that is 50 rupees.

(Refer Slide Time: 03:04)

Month	Interest to be paid for the month	Total Interest to be paid upto this month	Total Money owed upto this month
March	₹0.00	₹0.00	₹10,000.00
April	₹50.00	₹50.00	₹10,000.00
May	₹50.00	₹50.00	₹10,000.00
June	₹50.00	₹50.00	₹10,000.00
July	₹50.00	₹50.00	₹10,000.00
August	₹50.00	₹50.00	₹10,000.00
September	₹50.00	₹50.00	₹10,000.00
October	₹50.00	₹50.00	₹10,000.00
November	₹50.00	₹50.00	₹10,000.00
December	₹50.00	₹50.00	₹10,000.00
January	₹50.00	₹50.00	₹10,000.00

But we can do something more interesting when we go to the total interest to be paid up to this month. The total interest to be paid up to this month as you can realize is the interest for that month plus the total interest of the previous month. So, that would be the cell B3 + the cell C2.

(Refer Slide Time: 03:23)

Month	Interest to be paid for the month	Total Interest to be paid upto this month	Total Money owed upto this month
March	₹0.00	₹0.00	₹10,000.00
April	₹50.00	₹50.00	₹10,000.00
May	₹50.00	₹50.00	₹10,000.00
June	₹50.00	₹50.00	₹10,000.00
July	₹50.00	₹50.00	₹10,000.00
August	₹50.00	₹50.00	₹10,000.00
September	₹50.00	₹50.00	₹10,000.00
October	₹50.00	₹50.00	₹10,000.00
November	₹50.00	₹50.00	₹10,000.00
December	₹50.00	₹50.00	₹10,000.00
January	₹50.00	₹50.00	₹10,000.00

(Refer Slide Time: 03:36)

And this we can do by using again the “is equal to” symbol and we directly refer to the cell B3 as you can see the highlighted cell is B3, when I type in B3, the spreadsheet software highlights it plus C2 and that is also highlighted. And now I will press enter and I get the 50 rupees which is 50 plus 0.

(Refer Slide Time: 04:10)

Similarly, the total money owed up to this month would be the principal plus the total interest to be paid up to this month. So, I could again use cell reference and I could write this cell D3 to be equal to D2 + C3 and I get this.

Now, let us see what auto fill does here. I take this 50 and I drag it down like this and we see that the increments are all correct. 100 is this 50 plus this 50 and 150 is this 50 plus this 100 likewise, 200 is this 50 plus this 150. All these values have been filled in correctly. And how is that happening? Let us observe the formulae. This cell we wrote it as $B3 + C2$, but now when we go down to the next cell, it has changed to $B4 + C3$ and this one it has changed to $B5 + C4$.

(Refer Slide Time: 05:08)

Month	Interest to be paid for the month	Total Interest to be paid upto this month	Total Money owed upto this month
March	₹0.00	₹0.00	₹10,000.00
April	₹50.00	₹50.00	₹10,050.00
May	₹50.00	₹100.00	
June	₹50.00	₹200.00	
July	₹50.00	₹250.00	
August	₹50.00	₹300.00	
September	₹50.00	₹350.00	
October	₹50.00	₹400.00	
November	₹50.00	₹450.00	
December	₹50.00	₹500.00	
January	₹50.00		

So, what auto fill is doing is it is correspondingly changing the cells in the formula as we have gone down vertically as we have drag down vertically, it is only changing the row numbers, it is not changing the column values B is B, C is C; but as you can see $B6 + C5$, now it becomes $B7 + C6$. Now its becomes $B8 + C7$ and so on. So, this is giving us the values we want.

However, if we looked at the D column and we try to do the same thing, let us hold this and now drag it down. We see that there is a problem here. What I want is actually the principal plus the interest up to that point which should say this cell should be 10,150 whereas, this is 10,300.

This cell should be $10,000 + 200$, but it is now 10,500. So, what is going wrong is my formula is being updated to $D3 + C4$; the next one becomes $D4 + C5$. The next one becomes $D5 + C6$. But what we really want is the D2 to be fixed, the formula should not change D2; it should only change the C column row number.

(Refer Slide Time: 06:38)

Month	Interest to be paid for the month	Total Interest to be paid upto this month	Total Money owed upto this month
March	₹0.00	₹0.00	₹10,000.00
April	₹50.00	₹50.00	₹10,050.00
May	₹50.00	₹100.00	₹10,100.00
June	₹50.00	₹150.00	₹10,300.00
July	₹50.00	₹200.00	₹10,500.00
August	₹50.00	₹250.00	₹10,750.00
September	₹50.00	₹300.00	₹11,050.00
October	₹50.00	₹350.00	₹11,400.00
November	₹50.00	₹400.00	₹11,800.00
December	₹50.00	₹450.00	₹12,250.00
January	₹50.00	₹500.00	₹12,750.00
13			
14			
15			
16			
17			
18			
19			

And this can be done by putting a dollar symbol in the formula in front of the 2 and press enter. This cell does not change, but what the dollar symbol does when we do an auto fill now is, it keeps the cell fixed. And now, if you look at the formulae for the cells here, this is D2 + C4 now, this is D2 + C5, this is D2 + C6. In all cases we are getting the principal plus the corresponding interest up to that month.

Now, let us do something similar with this column. All of these instead of using the number 10,000; let us use cell reference. So, I will make this is equal to D\$2 once again because I only want the 10,000 amount and not the corresponding cells below. This value into again the same 0.5 by 100 enter and now again use what auto fill to bring this down and we see that, every cell is currently $D2 * 0.5 / 100$.

(Refer Slide Time: 08:11)

Month	Interest to be paid for the month	Total Interest to be paid upto this month	Total Money owed upto this month
March	₹0.00	₹0.00	₹20,000.00
April	₹100.00	₹100.00	₹20,100.00
May	₹100.00	₹200.00	₹20,200.00
June	₹100.00	₹300.00	₹20,300.00
July	₹100.00	₹400.00	₹20,400.00
August	₹100.00	₹500.00	₹20,500.00
September	₹100.00	₹600.00	₹20,600.00
October	₹100.00	₹700.00	₹20,700.00
November	₹100.00	₹800.00	₹20,800.00
December	₹100.00	₹900.00	₹20,900.00
January	₹100.00	₹1,000.00	₹21,000.00

And the reason this is useful, the reason cells reference is very useful is this. If I change my principal amount instead of 10,000 suppose I had made it 20,000; all the corresponding cells are automatically updated and this is because we are using cell reference. Every cell here depends on the 20,000, every cell here depends on the 20,000 and this one depends on the B column which is dependent on the 20,000 cell.

So, just this one change can give us an automatic update of all the month's information. So, you could give some strange number I mean an uncomfortable number like say 34,578 rupees and all the calculations are done for you.

(Refer Slide Time: 08:46)

Month	Interest to be paid for the month	Total Interest to be paid upto this month	Total Money owed upto this month
March	₹0.00	₹0.00	₹34,573.00
April	₹172.89	₹172.89	₹34,750.89
May	₹172.89	₹345.78	₹34,923.78
June	₹172.89	₹518.67	₹35,096.67
July	₹172.89	₹691.56	₹35,269.56
August	₹172.89	₹864.45	₹35,442.45
September	₹172.89	₹1,037.34	₹35,615.34
October	₹172.89	₹1,210.23	₹35,788.23
November	₹172.89	₹1,383.12	₹35,961.12
December	₹172.89	₹1,556.01	₹36,134.01
January	₹172.89	₹1,728.90	₹36,306.90
13			
14			
15			
16			
17			
18			
19			

Now, let us go back to making this 10,000 for looking at better numbers and this column which is currently doing the calculation with D2, but with a fixed interest rate that is 0.5 percent. We can make this column also a little better by using formulae, for that what we will do is, we will give the rate of interest in a particular cell.

(Refer Slide Time: 09:24)

Month	Interest to be paid for the month	Total Interest to be paid upto this month	Total Money owed upto this month	Interest rate per month >
March	₹0.00	₹0.00	₹10,000.00	0.5%
April	₹50.00	₹50.00	₹10,050.00	
May	₹50.00	₹100.00	₹10,100.00	
June	₹50.00	₹150.00	₹10,150.00	
July	₹50.00	₹200.00	₹10,200.00	
August	₹50.00	₹250.00	₹10,250.00	
September	₹50.00	₹300.00	₹10,300.00	
October	₹50.00	₹350.00	₹10,350.00	
November	₹50.00	₹400.00	₹10,400.00	
December	₹50.00	₹450.00	₹10,450.00	
January	₹50.00	₹500.00	₹10,500.00	
13				
14				
15				
16				
17				
18				
19				

In this cell, I will write down the title which is interest rate per month, enter and I would like to format it. Once again I will do text wrapping and now I will center align it a bit horizontal align, vertical align and maybe also give it a nice color, let us use this and

bold. So, this to indicate that it is the next cell, I will also give a small arrow looking symbol and the interest rate I am going to fill it in the cell which is our 0.5 percent and once again center align and done.

Now, what we could do is since we know that G2 in fact, I could just use the same formatting for this cell so, a very nice way to do that is I can use this cell, copy which is control C and then go here and paste control V and now change the content. So, I have managed to paste the format change the content to 0.5% and if we go back to our column here, the B column; we are going to refer to the cell G1 in order to calculate this value.

(Refer Slide Time: 11:07)

	D	E	F	G	H	I
1	GT	Total Money owed upto this month	Interest rate per month ->	0.5%		
2	GCD	₹10,000.00				
3	GTE	₹10,050.00				
4	GAMMA	₹10,100.00				
5	GAUSS	₹10,150.00				
6	GESTEP	₹10,200.00				
7	GROWTH	₹10,250.00				
8	GAMMALN	₹10,300.00				
9	GEOMEAN	₹10,350.00				
10	GAMMAINV	₹10,400.00				
11	INTERVENUE	₹10,450.00				
12	December	₹10,500.00				
13	January	₹10,500.00				

So, instead of $D2 * 0.5/100$, I am just going to write it as $D2 * G\$1$. In fact, if I want to specifically mention just that cell, I could also put a dollar in front of the column title which is G. So, I have D2 into dollar G dollar 1 enter. And now if I did the auto fill here, every cell is the same quantity which is the principal into the rate of interest.

(Refer Slide Time: 11:49)

The screenshot shows a Google Sheets spreadsheet titled "My First Spreadsheet". The data is as follows:

	A	C	D	E	F	G	H	I
1	Mar	Interest to be paid this month	Total Money owed upto this month		Interest rate per month ->	0.5%		
2	₹10.00	₹10,000.00			Principal ->	₹10,000.00		
3	₹50.00	₹10,050.00						
4	₹100.00	₹10,100.00						

A context menu is open over the cell containing "0.5%". The menu is titled "Paste special" and includes the following options:

- Paste values only
- Paste format only
- Paste all except borders
- Paste column widths only
- Paste formula only
- Paste data validation only
- Paste conditional formatting only
- Paste transposed

A green annotation on the right side of the menu reads: "Paste Special allows pasting particular properties of the cell".

I could go further and I can do the same thing with the principal. So, I am going to also enter the principal here and this principle, we are currently calling it 10,000 rupees, I could copy paste again, control C and control V.

And now if I wanted the same formatting as this, control C, I select these two cells and there is something called paste special which is we can go into the edit and in paste, control V was paste; but control alt V which is a special kind of paste, it paste only the format without the cell value. So, I just do this and now I have them all in the same format. Of course, this one can again be made number - Indian rupee.

(Refer Slide Time: 12:46)

Now, the convenience of all of this is, if I correspondingly changed all the mentions of a principal to that particular cell which is equal to \$G\$2. And this one also instead of D2, I will just refer to the cell with the principle which is \$G\$2. And thus, bring it all down and here also, I do not want it to refer to D2 any longer, I just want to G2 and again drag it all the way down.

(Refer Slide Time: 13:53)

Now, the convenience that we have with this kind of a setup is, you could automatically update the whole sheet simply by playing with these numbers. You want to change this

number to 15,000; you get the corresponding values and you want to change the interest rate, you want to make it 1% instead of 0.5; you getting the corresponding values, you could change it to 0.3% and the calculation is done.

So, this way you do not really need to input every cell of your spreadsheet. If you organize your spreadsheet well and you make sure it represents the data properly, you can also play around with your spreadsheet and update it in very easy ways. And in our example, the entire simple interest data for so and so number of months is completely dependent on these two cells G1 and G2. So, in this way organizing your data and building your spreadsheet in a way you can manipulate all your data conveniently is a very useful skill to have.

In the next video, we will see about some functions that you can use in your spreadsheets.

Thank you.

Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 10
Downloading and Uploading Spreadsheets

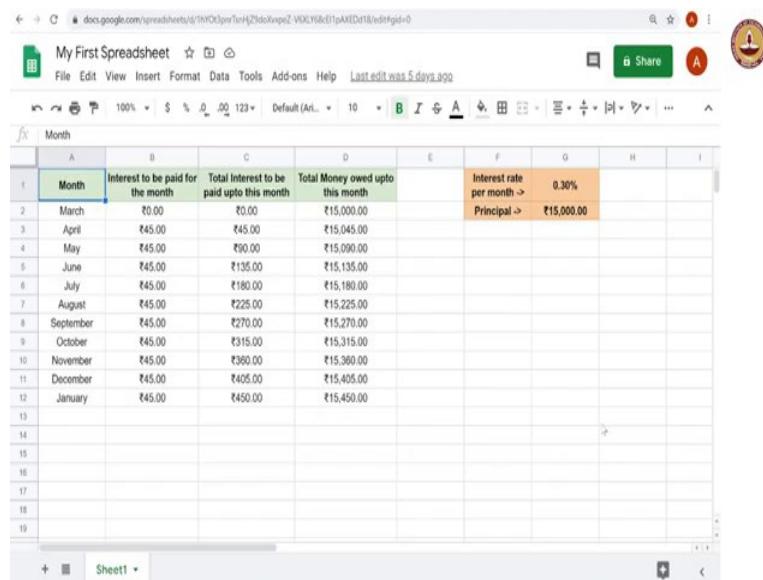
(Refer Slide Time: 00:13)

Learning Objectives for statistics week tutorials

Tutorial 4 Spreadsheets downloading and uploading Learning Objectives

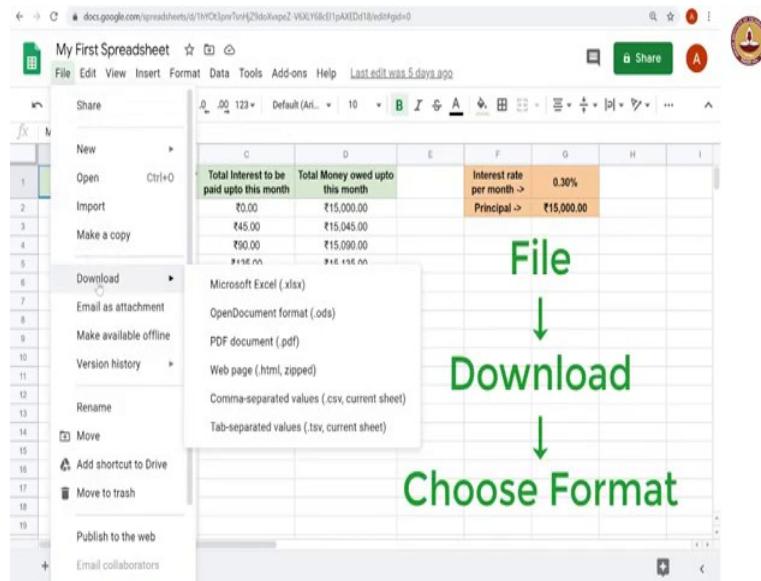
1. Downloading Google sheets file into a personal computer for offline use into various formats like XLSX, CSV, TSV, Pdf, ods.
2. Downloading CSV files from websites like data.gov.in.
3. Uploading the available offline files like CSV, XLSX files from the machine to Google Sheets.
4. Organizing the available google sheets in folders in Google Drive.

(Refer Slide Time: 00:16)



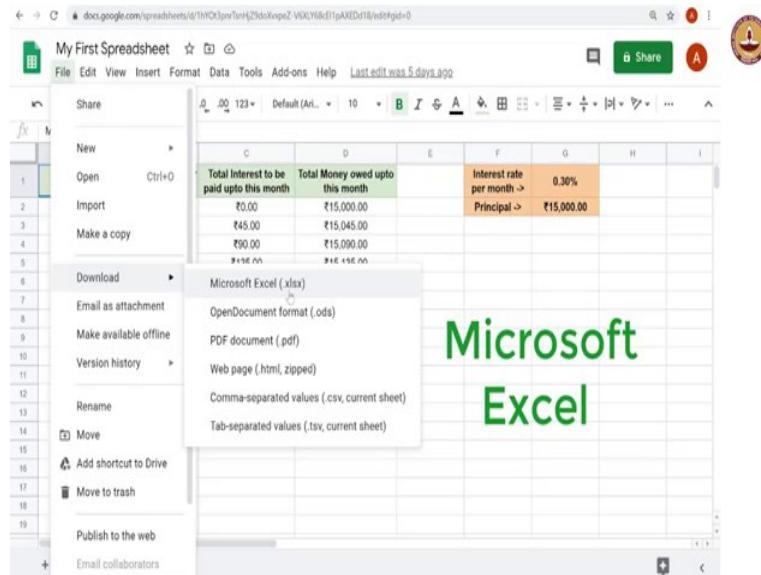
Month	Interest to be paid for the month	Total Interest to be paid upto this month	Total Money owed upto this month	Interest rate per month ->	Principal ->
March	₹0.00	₹0.00	₹15,000.00	0.30%	₹15,000.00
April	₹45.00	₹45.00	₹15,045.00		
May	₹45.00	₹90.00	₹15,090.00		
June	₹45.00	₹135.00	₹15,135.00		
July	₹45.00	₹180.00	₹15,180.00		
August	₹45.00	₹225.00	₹15,225.00		
September	₹45.00	₹270.00	₹15,270.00		
October	₹45.00	₹315.00	₹15,315.00		
November	₹45.00	₹360.00	₹15,360.00		
December	₹45.00	₹405.00	₹15,405.00		
January	₹45.00	₹450.00	₹15,450.00		
13					
14					
15					
16					
17					
18					
19					

(Refer Slide Time: 00:37)



Hello, welcome back. In this tutorial video, we will see how to save our spreadsheet from Google sheets; Google sheets is on the internet. You might want to download your spreadsheet into your machine into your computer. So, how do we do that is, we go to file and then we go to download and these are the various formats.

(Refer Slide Time: 00:49)



So, the most popular one is xlsx which is the format for a spreadsheet in Microsoft Excel which is a Microsoft product.

(Refer Slide Time: 00:59)

The screenshot shows a Google Sheets interface with a dropdown menu open from the 'File' menu. The menu includes options like 'New', 'Open', 'Import', 'Make a copy', 'Download' (with sub-options: Microsoft Excel (.xlsx), OpenDocument format (.ods), PDF document (.pdf), Web page (.html, zipped), Comma-separated values (.csv, current sheet), and Tab-separated values (.tsv, current sheet)), 'Rename', 'Move', 'Add shortcut to Drive', 'Move to trash', 'Publish to the web', and 'Email collaborators'. The main spreadsheet area displays a table with financial data:

Total Interest to be paid upto this month	Total Money owed upto this month	Interest rate per month ->	Principal ->
₹0.00	₹15,000.00	0.36%	₹15,000.00
₹45.00	₹15,045.00		
₹90.00	₹15,090.00		
₹135.00	₹15,135.00		

On the right side of the screen, there is a watermark or logo for 'Open Source Softwares Open Office/ Libre Office'.

And you could also download in the .ods format which is for Open Office or you could also use it for Libre Office. These are open source spreadsheet software and pdf is the portable document format which is not a spreadsheet, it will let you see the spreadsheet, but you cannot really treat it as a spreadsheet, you cannot open it again on Google sheets or Microsoft Excel or Libre Office right.

Web page is also similar to that, it is a dot html. You could open it on your browser such as Chrome or Firefox or Internet Explorer or whatever it is and this comma separated values which is the .csv file is something that can be used across platforms. For example, .xlsx works best with Excel, .ods works best with Open Office or Libre Office and .csv will open with whether it is Google sheets or Excel or Open Office.

(Refer Slide Time: 01:53)

The screenshot shows a Google Sheets document titled "My First Spreadsheet". A context menu is open under the "File" tab, specifically at the "Download" option. The menu lists several options: Microsoft Excel (.xlsx), OpenDocument format (.ods), PDF document (.pdf), Web page (.html, zipped), Comma-separated values (.csv, current sheet), and Tab-separated values (.tsv, current sheet). A green text overlay on the right side of the sheet reads "CSV files can be opened with multiple softwares".

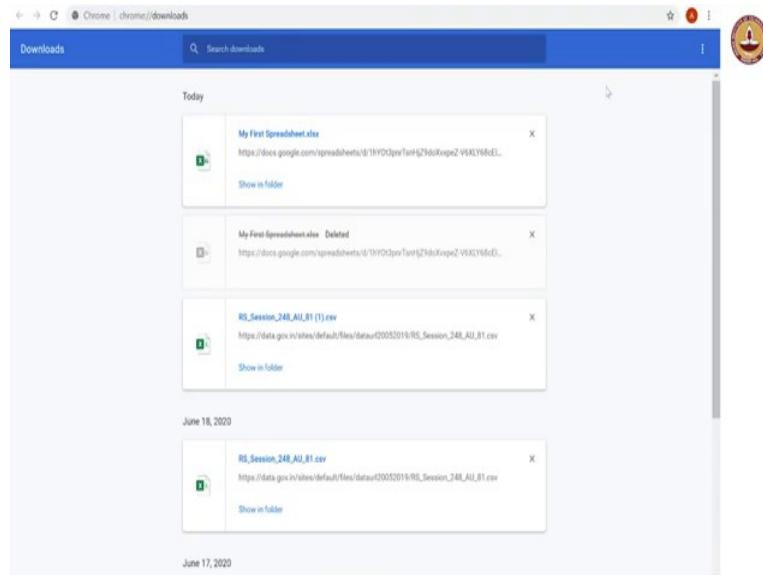
(Refer Slide Time: 02:12)

The screenshot shows a Windows File Explorer window with the path "Stats1 tutorials > Downloading and uploading spreadsheets". A save dialog box is open, prompting to save "My First Spreadsheet" as an ".xlsx" file. The "Save as type" dropdown is set to "Excel File (.xlsx)". In the background, a Microsoft Excel spreadsheet is visible, showing columns C, D, F, G, H, and I with some numerical data.

Anyway, for our purposes, let us download it as a Microsoft Excel file which is .xlsx. So, I click on that and now we have my Downloads folder showing. I could download it here or I could go back to This PC, these are all the folders that are there in This PC. I could go to Documents which is where I would like to put my Stats1 tutorials. So, this is another folder and within this folder, I will create a new folder which happens by right click and then go to New and then Folder. This one I will call "Downloading and uploading spreadsheets".

Once I type that press enter. Now open this folder and here I will save my spreadsheet file. Chrome is showing that it is been downloaded.

(Refer Slide Time: 03:22)



Now, how do I go there? I could go to downloads and in downloads, I can look at show in folder and here we have our Excel file. I can open this on Excel. Now let us close this and we will see what to do with downloading data sets from the internet.

(Refer Slide Time: 03:58)



So, let us go to the internet. A good source for downloading data sets is data.gov.in which has a lot of India-related based data sets, it is the government website. So, you have all these options.

(Refer Slide Time: 04:23)

The screenshot shows a search results page for 'traffic' on the data.gov.in website. The results are as follows:

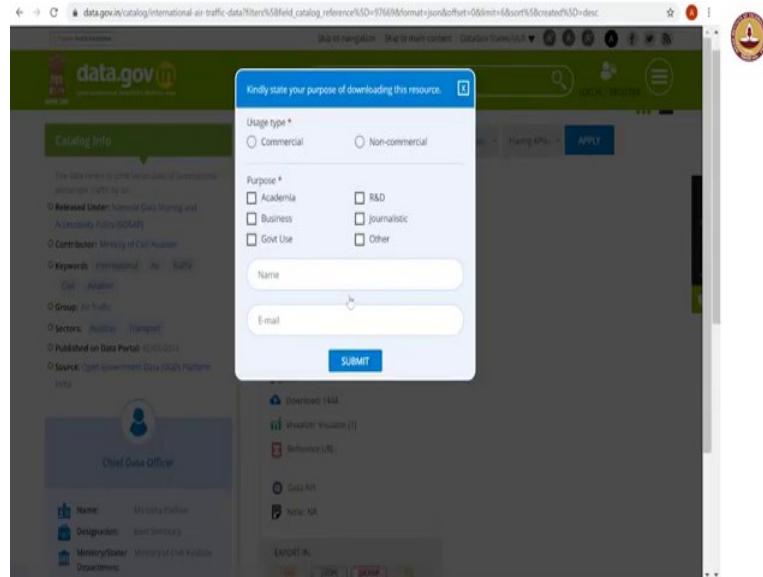
- Railway Traffic And Earnings**: Get Railway Traffic and Earnings data of India. Data provides Railway Traffic and Earnings from Goods Traffic... [+]
- Quarterly Air Traffic Statistics**: The catalog Quarterly Air Traffic Statistics contains Quarter wise various operating and traffic... [+]
- Monthly Air Traffic Statistics**: The catalog Monthly Air Traffic Statistics contains Month wise various operating and traffic... [+]
- International Air Traffic Data**: International Air Traffic Data
The data refers to time series data of international passenger traffic by air.
International Air Traffic Data
- Telecom Traffic Pattern - CDMA**: The data refers to the circle wise detail of traffic pattern of CDMA.

Let us say we will look at something on traffic and now you see all these options. Let us say we look at International Air Traffic Data, I clicked on that and it is a csv file.

(Refer Slide Time: 04:37)

The screenshot shows the details of the 'International Air Traffic Data' dataset on the data.gov.in website. The dataset is a CSV file, 1.4 KB in size, with 1444 rows. It was published on 02/05/2014. The dataset is categorized under International Air Traffic and is available in CSV format. The dataset is described as having annual granularity. The dataset is contributed by the Ministry of Civil Aviation and is released under the National Data Sharing and Accessibility Policy (NDSAP). The dataset is also available as a Data API.

(Refer Slide Time: 04:40)



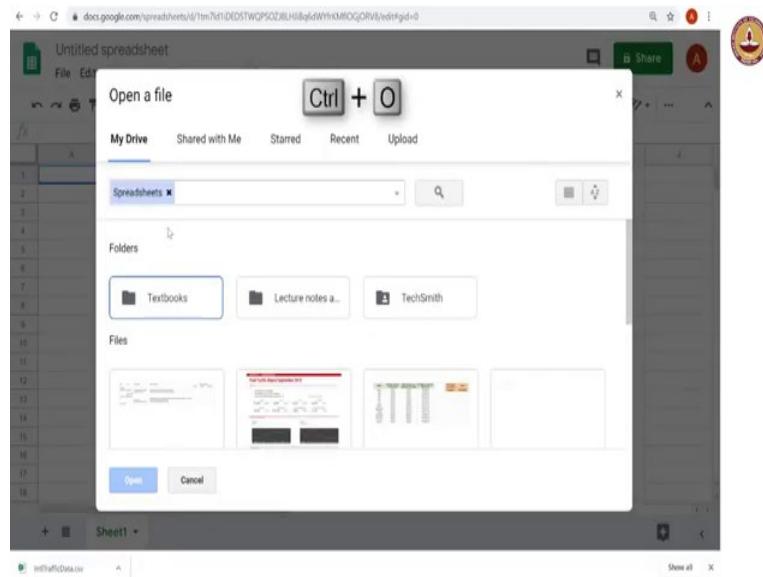
So, I will click on that too, it asks me these details; I will fill them up and once I have filled them up, begins to download as a csv file; save. Now I would like to open this csv file in Google sheets, the csv file I downloaded from data.gov.in.

(Refer Slide Time: 05:14)

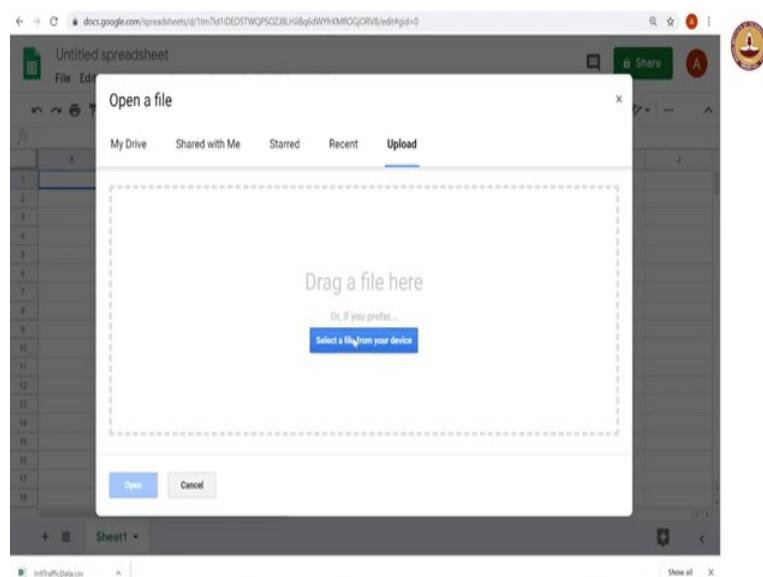
A screenshot of a Google Sheets spreadsheet titled 'IntlTrafficData'. The spreadsheet contains data from 2005 to 2011. The columns represent the year, operator, Pax To, Pax From, Pax Total, Freight To, Freight Fwd, Freight Total, and other related metrics. The data is presented in a tabular format with rows for each year and individual entries for each operator and route.

So, let me go back to the spreadsheet and to go to the spreadsheet, home I click on this sheets home icon there and here we are. Now, let us open a new spreadsheet, blank and here let us open our csv file. For this what we do? File and Open; control O and I would like to upload my file from the machine.

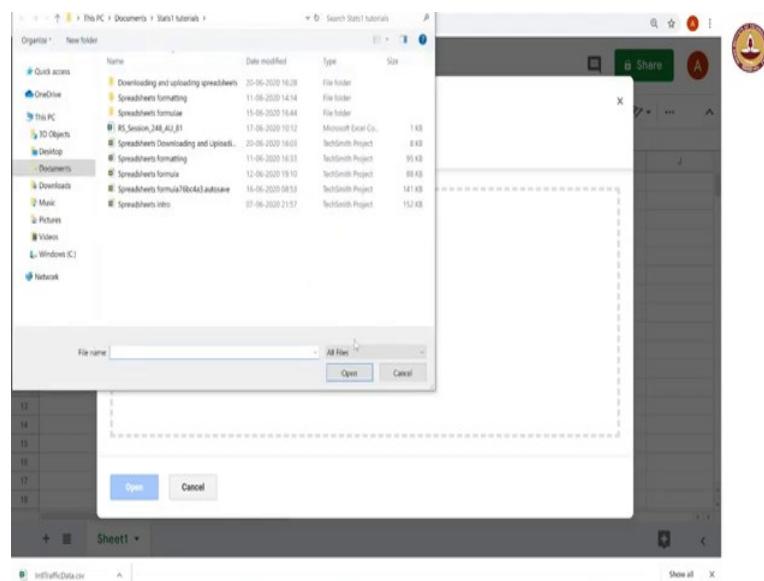
(Refer Slide Time: 05:29)



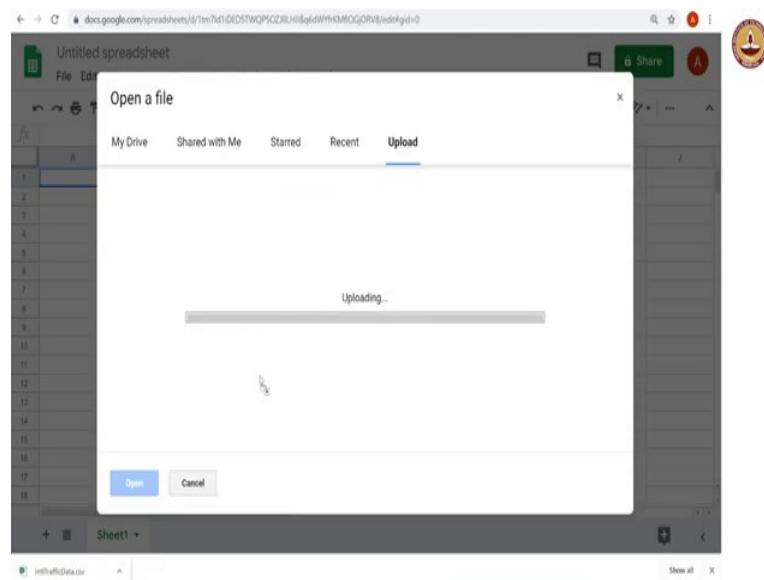
(Refer Slide Time: 05:38)



(Refer Slide Time: 05:41)

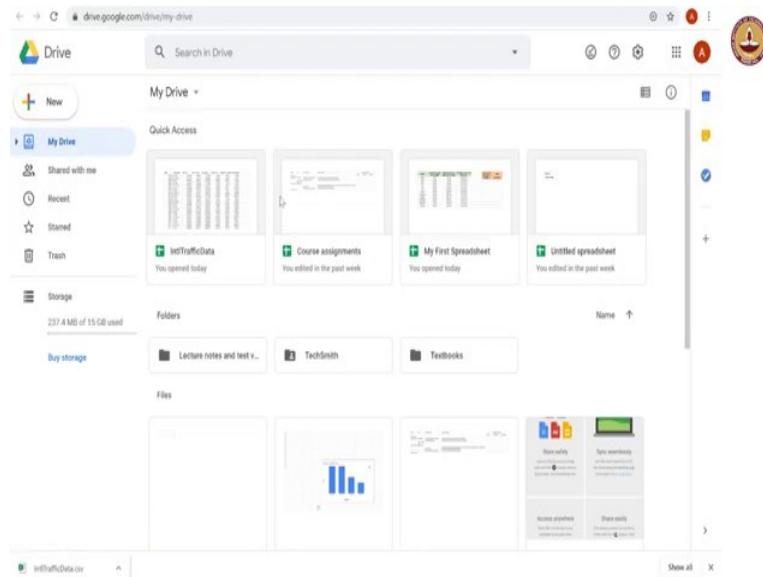


(Refer Slide Time: 05:49)



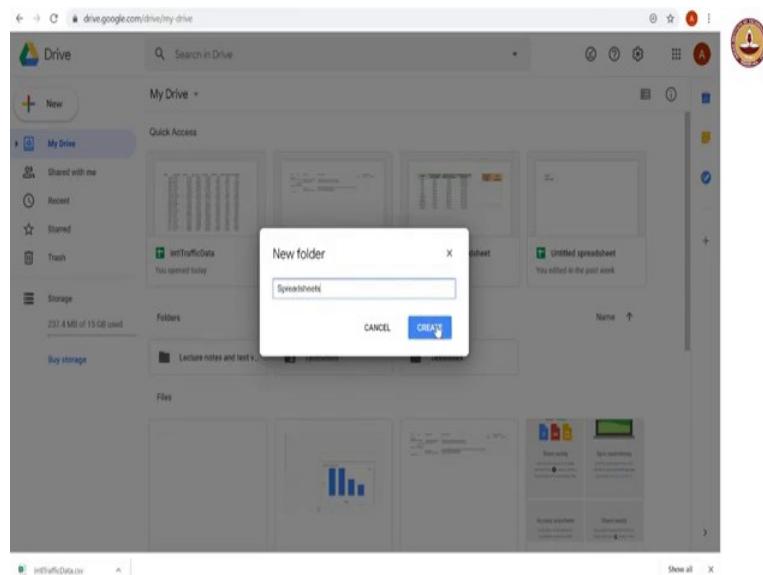
So, I will select a file from the device or I could simply just drag this, I could click hold and drag it here and it opens up and this is the data that we have downloaded from the internet. So, this way you could download and use datasets that you have gotten from the internet.

(Refer Slide Time: 06:20)



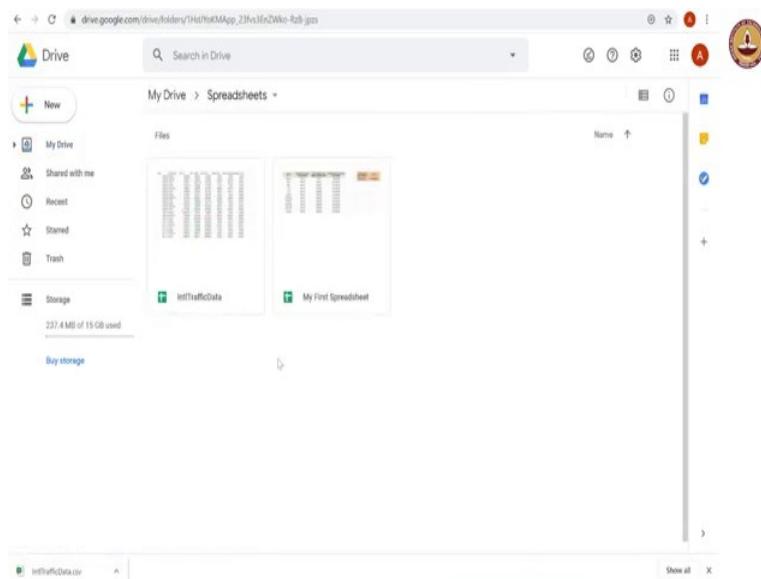
And now how do we arrange all these sheets? So, let us go back to sheets home. Now, let us go to drive - Google drive which in this option we have it here drive, drive is arranged like the folder system that we have seen on our machine. So, the folder system as you can see these are the folders here. Let us open any of those and I could go back to My Drive and here we are. If we went to this folder, again I could go back to My Drive in this way and I might want to keep all these spreadsheets in a folder.

(Refer Slide Time: 06:58)



So, I can create a folder by right clicking adding a New folder; I will call it simply spreadsheets, I created and into this folder, I would like to add these spreadsheets. For example, this international traffic data that we just downloaded, I can click hold and drag into this folder, likewise our spreadsheet example on simple interest. I can also again hold this drag and put in the spreadsheets folder.

(Refer Slide Time: 07:38)



So, now if I open the spreadsheets folder, I have these two spreadsheets inside it. So, this way you can organize your drive. You could just the way you can organize the folders and files in your computer and whatever you have in your Google sheets, you could directly open from there and begin to work with the data.

Thank you.

Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture No. 2.1
Describing Categorical Data – Frequency distributions

In the last lecture, you were introduced to basically the two important branches of statistics, which are descriptive statistics and inferential statistics. You also understood what is the difference between a sample and population.

(Refer Slide Time: 00:32)

The screenshot shows a presentation slide with the title 'Statistics for Data Science - 1' at the top left. To the right is the Indian Institute of Technology Madras logo. Below the title, the word 'Review' is written in blue. A list of three points follows:

1. What is statistics?
 - ▶ Descriptive statistics, inferential statistics.
 - ▶ Distinguish between a sample and a population.
2. Understand how data are collected.
 - ▶ Identify variables and cases (observations) in a data set
3. Types of data-
 - ▶ classify data as categorical(qualitative) or numerical(quantitative) data.
 - ▶ Understand cross-sectional versus time-series data.
 - ▶ Measurement scales

A photograph of Prof. Usha Mohan is visible on the right side of the slide, and a navigation bar with icons is at the bottom.

But we restricted our introduction to what is a sample and population for what is required for this particular course. Of course, in the advanced courses you will learn more about sample and population. Then we went on to identify and understand what is a data. Again for this purpose of this course we are restricting ourselves to only structured data.

In the structured data, I have data in the form of a table where the variables are recorded in a column and the observations are cases which are represented by the rows in a data set. We looked at the types of data, broadly you classify data as categorical data and numerical data. We also understood what is the difference between cross-sectional data and time series data.

And finally, you should be knowing by this time, what are the measurement scales by that I mean that when you have categorical data, you should know how to distinguish between whether

it is a nominal variable or an ordinal variable and when it comes to numerical data, you should know whether it is an interval variable or ratio variable. This is what you should be knowing at this point of time.

(Refer Slide Time: 01:50)

The screenshot shows a video player interface. At the top left, it says 'Statistics for Data Science -1'. In the top right corner is the logo of the Indian Institute of Technology Madras. The main content area displays the title 'Statistics for Data Science -1' and a subtitle 'Describing Categorical Data- Single Variable'. Below this, the name 'Usha Mohan' is shown, followed by the text 'Indian Institute of Technology Madras'. A video frame shows a woman with glasses and an orange sari speaking. At the bottom of the video frame are standard video control icons for navigation and volume.

Moving forward, you are going to understand how to describe categorical data. Now, again in this module you are going to understand first, we start with describing categorical data for a single variable, and then we will look at measures of association when they have more than one variable. So, today what we are going to do is, we are going to understand how to describe categorical data.

(Refer Slide Time: 2:13)



Frequency distributions
Relative frequency distributions

Charts of categorical data

- Pie charts
- Bar charts
- Pareto charts

Mode and Median



Navigation icons: back, forward, search, etc.

So, we start with what we understand as a frequency distribution.

(Refer Slide Time: 02:20)



Frequency distributions

Definition

A *frequency distribution*¹ of qualitative data is a listing of the distinct values and their frequencies.

Each row of a frequency table lists a category along with the number of cases in this category.

COUNT

¹Weiss, Neil A. Introductory Statistics: Pearson New International Edition.
Pearson Education Limited, 2014.



Navigation icons: back, forward, search, etc.

Now, the definition of frequency definition is a listing of distinct values and their frequencies. What do we mean by frequency? Frequency is nothing but the count, nothing but count. And by distinct values, you mean what are the distinct values the categorical variable actually takes. This is what we mean by that. Now, each row of a frequency table lists a category along with the number of cases or count of cases. The minute I say the number of cases, I am just, I imply how many of that particular case is there in that particular category.

(Refer Slide Time: 03:14)

Statistics for Data Science -1
↳ Frequency distributions

Example

Construct a frequency table for the given data

→ 1. A,A,B,C,A,D,A,B,D,C
2. A,A,B,C,A,D,A,B,D,C,A,B,C,D,A
3. A,A,B,C,A,A,B,B,D,C,A,B,C,D,B
4. A,A,B,C,A,D,A,B,D,C, A,B,C,D,A,C,D,D

Category	
A	
B	
C	
D	

The slide also features a woman in an orange sari speaking.

So, for example, let's look at constructing frequency table for this simple given data. The category is just alphabets. I have the alphabets, I have 4 categories, I can term them as A, B, C, and D.

(Refer Slide Time: 3:34)

Statistics for Data Science -1
↳ Frequency distributions

Construct a frequency distribution

The steps to construct a frequency distribution²

Step 1 List the distinct values of the observations in the data set in the first column of a table.

Step 2 For each observation, place a tally mark in the second column of the table in the row of the appropriate distinct value.

Step 3 Count the tallies for each distinct value and record the totals in the third column of the table.

The slide also features a woman in an orange sari speaking.

²Weiss, Neil A. Introductory Statistics. Pearson New International Edition.
Pearson Education Limited, 2014.

So, how do I construct a frequency distribution? List the distinct values of observation. Here, what are the distinct values of my observation? In the first example, my distinct values of observation are A, B, C, and D. I list them and I write it as a category. This is the first thing I do. That is my step one. Step one is to list the distinct values and this is my first column.

(Refer Slide Time: 4:06)

Statistics for Data Science -1
↳ Frequency distributions

Example

1. A,A,B,C,A,D,A,B,D,C

Category	Tally mark	Frequency
A		4
B		2
C	/	1
D		2
Total	10	10

2. A,A,B,C,A,D,A,B,D,C, A,B,C,D,A

Category	Tally mark	Frequency
A		6
B		3
C		2
D		3
Total	15	15



So you can see that in the first column, I listed the distinct values which are A, B, C, and D. For each observation, place a tally mark in the second column. So for A, I place a tally mark, which I am marking here, the second observation again is A, I again have a tally mark, third observation is B, fourth observation is a C, fifth observation is again A, sixth observation is a D, seventh observation is an A, then I have a B, then I have a D, I have a C.

So, these are the tally marks which I am going to have. So, that is step two. For each observation, place a tally mark and then count the number of tallies. So, if I count the number of tallies, the frequency, or the count of A is 4, count of B is 2, count of C is 2, and count of D is 2. This is what I refer to as a frequency distribution, where the distinct values are given in column 1, tally marks in column 2, and the count in column 3.

Now let's go and look at this example. If you look at this example, again I have a tally mark. I will do it quietly, faster at this time, A, A, B, C, A, D, A, B, D, C. Now this A, I just cross it out, because this is my fifth value. Whenever I have more than 4, I am having a tally mark. I am crossing it out as the fifth value, then I have a B, C, D, and A. So, this $5 + 1$ gives me a value of 6, this is a 3, this is a 3, this is a 3.

So, I have a total of 15 observations in this case, where this is the distribution, category A occurs six times, B occurs 3 times, C occurs three times, and D occurs three times. So, if you look at

this data where A had a tally mark of 4, B had 2, C had 2, and D had 2 with a total 10. This was a 4, 2, 2, 2. This is a 6, 3, 3, 3.

(Refer Slide Time: 06:40)

Statistics for Data Science -1
↳ Frequency distributions

Example

3. A,B,B,C,A,D,B,B,D,C, A,B,C,D,B

Category	Tally mark	Frequency
A		3
B		6
C		3
D		3
Total		15



Now, let's look at this here again. The same data of 15 points, but now I have a A, I have a B, B, I have a C, I have a A, I have a D. I have a B, B, D, C, a A, a B, a C, a D, a B. So, this has A is appearing 3 times, B is appearing 6 times, C is appearing 3, and D is appearing 3. I have again a total of 15 observations. So, if you look at compare this example with the earlier example, you see that the only difference between example 2 and example 3 is, A appear 6 times here and B appears three times, C and D appear 3 times each.

In this example, it has flipped between A and B with A appearing thrice, B appearing 6 times, and C and D appearing 3 times each. Both of them have 15 observations, and this is what is given.

(Refer Slide Time: 7:48)



Example

3. A,B,B,C,A,D,B,B,D,C, A,B,C,D,B

Category	Tally mark	Frequency
A		3
B		6
C		3
D		3
Total		

4. A,A,B,C,A,D,A,B,D,C, A,B,C,D,A,C,D,D

Category	Tally mark	Frequency
A		6
B		3
C		4
D		5
Total		18



Now, let's look at a final example where I have so many observations. Again if I do a tally mark A, A, B, C, A, D, A, B, D, C, A, B, C, D, A, C, D, D. So if I look at this, I have a 6, I have a 3, I have a 4, I have a 5. So, I have 18 observations here. Now, if you look at this, I have a tally mark here. I had a 3, I had a 6, I had a, then I had 1, 2, 3, I had a 3. This is what I have here. So, you can see that we can construct different frequency distributions.

(Refer Slide Time: 8:48)

Frequency table in a googlesheet

- Step 1 Select/Highlight the cells having data you want to visualize.
- Step 2 In the Formatting bar click on the Data option.
- Step 3 In the Data option go to Pivot Table option and create a new sheet.
- Step 4 After creating Pivot Table go in Pivot Table Editor and in that first add rows and then values.



How do I do a frequency table in a Google Sheet? So, let's look at how to do a frequency table in a Google Sheet. I look at the same example I have taken.

(Refer Slide Time: 9:00)

CATEGORY	COUNT of CAT
A	4
B	2
C	2
D	2
Grand Total	10

CATEGORY	FREQUENCY
A	4
B	2
C	2
D	2
Grand Total	10

So this is the Google Sheet I have. I am going to construct, I am going to add a sheet here. In that sheet, I am going to type first the category name. So, I have a category here. I am going to write down whatever data I have here, the data, I am going to list on the data. I have a A, I have A, B, C, A, D, A, B, D, C. So, you can see that this is the first example which we look, I have listed down the data here.

So, you go back to the step one, select highlight the cells you have, what are the cells I have, I just have these cells, I am highlighting these cells. That is the first step. Now you look at the second step in the formatting bar. Click on the data option. So, I go to the formatting bar, I click on the data option. Then, in the data option, go to the pivot table option. Go, first I highlight my data, I go to the data option, I have what I call a pivot table option.

Now, this pivot table I specify the range. You see that the range is specified A one to A eleven with the cell A one specifying what is the category. Here I have just given the name category, I could give any name to this category variable. I could give an alphabet or I could just tell it is some group, anything, but this I am just specifying a certain category, asking it to create a pivot table.

Now, let us go to the final step. After creating the pivot table in the pivot table editor, what is the pivot table editor, you have the pivot table editor which appears on the right-hand side. There you add rows. In the row, I just add a category. What are the different categories? I have

category A, B, C, and D. And in the values, I'm going to add what are the values that category has which is 4, 2, 2, 2.

So, one way to create a table is, I can just copy this and I paste the values. I can give a category table here with frequency here and I can see that this is nothing but the table we have just created. So, this is one way to create a frequency table in your Google Sheet.

So, once you have your frequency table, your frequency table, so this is precisely what we did for the first example 4, 2, 2, 10. You can see that is what our Google Sheet gives, A frequency 4, B frequency 2, C frequency 2, and D frequency 2 with a grand total of 10. This is what we have here, the first frequency table which you have created on a Google Sheet.

(Refer Slide Time: 12:49)

S.No	Date(dd/mm/yyyy)	Time (IST)	Height (cm)	Gender	Weight (Kg)	Blood Group	Body Temperature (F)	Blood Pressure
1	2/3/2020	7:30:00 AM	178	M	75	O+	100	118/80
2	2/3/2020	8:00:00 AM	150	F	57.5	A-	98.4	125/85
3	2/3/2020	8:12:00 AM	162	M	61	O-	98.2	120/80
4	2/3/2020	8:52:00 AM	145	M	65	B+	78.5	123/82
5	2/3/2020	9:00:00 AM	153	M	72	A+	95.5	109/88
6	2/3/2020	9:09:00 AM	167	M	98	O+	110	155/95
7	2/3/2020	10:00:00 AM	175	M	69	B-	94	116/80
8	2/3/2020	10:10:00 AM	165	F	59	O-	93	115/80
9	3/3/2020	7:40:00 AM	169	M	65	A+	98	130/85
10	3/3/2020	7:59:00 AM	173	M	74	AB+	101.1	130/83
11	3/3/2020	8:01:00 AM	156	M	61	O-	98.9	126/82
12	3/3/2020	8:15:00 AM	158	F	52	B+	99.7	135/85
13	3/3/2020	8:41:00 AM	183	M	82	AB-	102	123/82
14	3/3/2020	9:00:00 AM	167	M	71	B-	90.9	134/89
15	3/3/2020	9:30:00 AM	169	M	63	A+	94.5	118/79
16	4/3/2020	7:20:00 AM	171	M	70	AB+	97.5	115/76
17	4/3/2020	8:27:00 AM	163	F	67	O-	94	121/83
18	4/3/2020	9:45:00 AM	155	F	64	B-	95.7	115/75
19	4/3/2020	9:58:00 AM	150	M	55	A+	100	117/77
20	4/3/2020	8:39:00 AM	145	F	58	AB-	94.8	122/83
21	4/3/2020	7:55:00 AM	174	M	74	O+	99	127/88
22	4/3/2020	10:18:00 AM	167	M	68	B+	101	114/72
23	4/3/2020	9:37:00 AM	162	F	71	A+	94.6	119/78
24	5/3/2020	7:36:00 AM	158	F	56	O+	99	128/87
25	5/3/2020	10:45:00 AM	149	F	49	B+	98.7	118/79

Now again, I can create the same table for any given data. If you recall, this is the blood group data, hospital data which we discussed in our earlier class. Suppose I want to know the distribution of a particular categorical variable here, I can look at two categorical variables here. One is gender and other is blood group. When I look at blood group, this is what I have here. This is the categorical variable.

So, I go back to my pivot table step. So, what do I do, I remember, in first I select the cells, I click on the data option, I go to the pivot table option and I go to the pivot table editor. So, we

are going to do the same thing here. I select this data, I click on data option. Here I click on the pivot table option.

(Refer Slide Time: 13:47

The screenshot shows a Google Sheets interface with a pivot table editor open. The main table (A1:L10) has columns for 'Blood Group' and 'COUNT of Blo'. The data rows show counts for A-, A+, AB-, AB+, B-, B+, O-, and O+. Row 10 is labeled 'Grand Total' with a value of 30. The second table (A13:L22) has columns for 'Blood Group' and 'FREQUENCY'. It contains the same data as the first table, with row 22 labeled 'Grand Total' with a value of 30. The pivot table editor toolbar at the bottom includes buttons for 'Sheet1', 'Pivot Table 1', 'Pivot Table 2', and 'Pivot Table 3'.

Blood Group	COUNT of Blo
A-	2
A+	6
AB-	3
AB+	2
B-	3
B+	4
O-	5
O+	5
Grand Total	30

Blood Group	FREQUENCY
A-	2
A+	6
AB-	3
AB+	2
B-	3
B+	4
O-	5
O+	5
Grand Total	30

I create a new sheet. Once I create a new sheet, I go to the pivot table editor. I add the rows. Now you can see the name of the variable, now here is a blood group. That is what I want to know the frequency distribution of the variable blood group. So, I click on blood group here, and I go to values and I add the values and you can see that this is the frequency distribution of the blood group. I can just copy, I can paste the values. And I can just put here, this is the blood group, and this is the frequency.

And you can see that this is the frequency distribution of table of my blood group, which I get in Google Sheets. If you look at the sum, you can see the sum of all of those study and that is precisely I have 30 observations, this is blood group with a frequency of 30 people. So, this is how we construct frequency tables both manually that is through first principles of using tally marks and this is through using a Google Sheet. Frequency table gives the count of each variable, each categorical variable.

(Refer Slide Time: 15:43)

Statistics for Data Science -1
└ Frequency distributions
 └ Relative frequency distributions



Relative frequency



Definition

The ratio of the frequency to the total number of observations is called **relative frequency**

- ▶ The steps to construct a relative frequency distribution
 - Step 1 Obtain a frequency distribution of the data.
 - Step 2 Divide each frequency by the total number of observations.

There is another thing which is very useful and that is called relative frequency. What relative frequency captures is the ratio of the frequency to the total number of observations.

(Refer Slide Time: 15:59)

Statistics for Data Science -1
└ Frequency distributions
 └ Relative frequency distributions



Example

1. A,A,B,C,A,D,A,B,D,C

Category	Tally mark	Frequency	Relative frequency
A		4	0.4
B		2	0.2
C		2	0.2
D		2	0.2
Total		10	1

2. A,A,B,C,A,D,A,B,D,C, A,B,C,D,A

Category	Tally mark	Frequency	Relative frequency
A		6	0.4
B		3	0.2
C		3	0.2
D		3	0.2
Total		15	1



So, we already have constructed a frequency table, we already saw that there are 4 A, 2 B, 2 C, and 2 D. The ratio of the frequency, so, 4 is the frequency of A, total number of observation is 10. The ratio that is 4/10, which is 0.4, gives me the relative frequency of A in this table. So,

relative frequency distribution, I just divide each frequency by the total number of observations, and I get a 0.4, 0.2, 0.2, 0.2. And the sum total of a relative frequency should always add up to 1.

Now, if you look at this data, this was a 0.4, 0.2, 0.2, and 0.2 and adding up to 1. Here you look at it, this is going to be $6/15$, this is going to be $3/15$, this is going to be a $3/15$, this is going to be a $3/15$. And you can see that all of them add up to $15/15$, which is 1. You can again see that this is a 0.4, this is 0.2, this is 0.2, this is 0.2.

Now, what I want you to see is, the frequency of these two distributions are different. Here, I had an A which is 4, 2, 2, 2. Here the frequency was 6, 3, 3, 3. Whereas when you look at the relative frequency of this dataset and this dataset, you can find that the relative frequency of A in this dataset, the relative frequency of each one of the categories or each one of the variables A, B, C, and D is the same as the relative frequency of A, B, C and D in this dataset.

(Refer Slide Time: 18:07)

The screenshot shows a presentation slide with the following elements:

- Navigation bar: Statistics for Data Science -1, L-Frequency distributions, L-Relative frequency distributions.
- Logo: Institute of Technology, RGPV logo.
- Title: Why relative frequency?
- Content:
 - For comparing two data sets.
 - Because relative frequencies always fall between 0 and 1, they provide a standard for comparison.
- Speaker: A woman in an orange sari speaking at a podium.
- Navigation icons: Back, forward, search, etc.

Now, the reason for why do we need relative frequency, as I have demonstrated here even though there is a difference between these two datasets in the count, you can see the relative frequency is pretty much the same. Here, I had totally 10 observations, here I had totally 15 observations, but both of them have the same relative frequency, the frequency is different.

So, what relative frequency helps us, is to compare two datasets. And because relative frequency always is between 0 and 1, it is a good standard for comparison. Hence, we always prefer to have a relative frequency table.

(Refer Slide Time: 19:06)

The screenshot shows a Google Sheets document titled "Hospital Data". It contains two tables. The first table, titled "Blood Group", has columns for "Blood Group" and "COUNT of Blo". The data includes rows for A-, A+, AB-, AB+, B-, B+, O-, and O+. The last row is a "Grand Total" with a value of 30. The second table, titled "FREQUENCY", has columns for "Blood Group", "FREQUENCY", and "RELATIVE FREQUENCY". The data is identical to the first table, with frequencies 2, 6, 3, 2, 3, 3, 4, 5, and 5 respectively, and a total of 30. The "RELATIVE FREQUENCY" column shows values like 0.06666666666666667, 0.2, 0.1, etc.

Blood Group	COUNT of Blo
A-	2
A+	6
AB-	3
AB+	2
B-	3
B+	4
O-	5
O+	5
Grand Total	30

Blood Group	FREQUENCY	RELATIVE FREQUENCY
A-	2	0.06666666666666667
A+	6	0.2
AB-	3	0.1
AB+	2	0.06666666666666667
B-	3	0.1
B+	4	0.1333333333333333
O-	5	0.16666666666666667
O+	5	0.16666666666666667
Grand Total	30	1

How do we create a relative frequency table in Google Sheet? In Google Sheet, we already have a frequency table. I create what is called a relative frequency column, here, and I know that relative frequency is nothing but the frequency divided by the total. That is how I define it. And I can just drag this down. And you can see that, if I look at the sum of these values, it adds up to 1 with each of these frequencies giving me the relative frequency. This is for the blood group.

(Refer Slide Time: 20:03)

The screenshot shows a Google Sheets document titled "Hospital Data". It contains two tables. The first table, titled "CATEGORY", has columns for "CATEGORY" and "COUNT of CAT". The data includes rows for A, B, C, and D. The last row is a "Grand Total" with a value of 10. The second table, titled "FREQUENCY", has columns for "CATEGORY", "FREQUENCY", and "RELATIV FREQ". The data is identical to the first table, with frequencies 4, 2, 2, and 2 respectively, and a total of 10. The "RELATIV FREQ" column shows values like 0.4, 0.2, 0.2, etc.

CATEGORY	COUNT of CAT
A	4
B	2
C	2
D	2
Grand Total	10

CATEGORY	FREQUENCY	RELATIV FREQ
A	4	0.4
B	2	0.2
C	2	0.2
D	2	0.2
Grand Total	10	1

I can do the same thing for this pivot table. I can do the relative frequency, which is equal to just a 0.4, 0.2, 0.2, 0.2, and the sum of all relative frequencies would always add up to 1. So, now I have this, which is going to give me, this is not the blood group, this is category. So, I have for the first example, I have with me what I call the frequency and the relative frequency listed along with the category variable.

(Refer Slide Time: 20:56)

Statistics for Data Science - 1
└ Frequency distributions
 └ Relative frequency distributions



Example

3. A,B,B,C,A,D,B,B,D,C, A,B,C,D,B

Category	Tally mark	Frequency	Relative frequency
A		3	0.2
B		6	0.4
C		3	0.2
D		3	0.2
Total		15	1

4. A,A,B,C,A,D,A,B,D,C, A,B,C,D,A,C,D,D

Category	Tally mark	Frequency	Relative frequency
A		6	6/18
B		3	3/18
C		4	4/18
D		5	5/18
Total		18	1.



So, I have two more examples. In the earlier examples, so, this is going to be 3/15, which is a 1 by 5, so I have a 0.2, 6 /15, which is 2/5, which is 0.4. I have again 0.2, I have 0.2. So, for each one of them, you can see that it adds up to 1. I leave this as an exercise, but you can see that this is going to be 6 /18, this is going to be 3/18, this is going to be 4/18, and 5/18. You can see that this adds up 5+4, 9, 9+ 9, 18. So this adds up again to 1.

(Refer Slide Time: 21:35)



Summary

1. Constructing a frequency table.
2. Notion of relative frequency and constructing a relative frequency table.



So in summary, what we have learned in this portion is how to construct a frequency table? What is the notion of relative frequency? And how do you construct a relative frequency table?

Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture No. 2.2
Describing Categorical Data – Charts of categorical data

(Refer Slide Time: 00:14)

Statistics for Data Science -1
└ Charts of categorical data



Charts of categorical data

- ▶ The two most common displays of a categorical variable are a bar chart and a pie chart.
- ▶ Both describe a categorical variable by displaying its frequency table.

Now when we come to graphical displays of categorical data what we have learned so far is how to come up with a frequency table and we just also demonstrated how to come up with a relative frequency table. So, once I have the tabular form of summarizing my data and remember we are only working with categorical data at this point of time, and I am looking at only one variable. The next common thing is how do I graphically display this data.

So, when it comes to categorical data, the two most common displays of a categorical data are a bar chart and a pie chart. Since we introduced already a frequency table, we also see that this both the pie chart and the bar chart basically display the data that is given in the frequency table. What do we mean by that?

(Refer Slide Time: 01:14)

Statistics for Data Science -1

└ Charts of categorical data

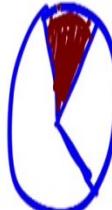
 └ Pie charts

Pie charts



Definition

WEDGES



A *pie chart* is a circle divided into pieces proportional to the relative frequencies of the qualitative data.

Recall that a pie chart. So, I can define a pie chart as a circle that is divided into pieces or wedges. Some textbooks refer to this as wedges. The reason why it is called a wedge is a pie chart is a circle or a disc which is divided into pieces or wedges. Now this portion I have here is a wedge, this particular portion is a wedge. Now again, another reason why it is called a pie chart is if this is a pie, this shaded portion is basically a share of a pie or a piece of a pie.

(Refer Slide Time: 2:10)

Statistics for Data Science -1

└ Charts of categorical data

 └ Pie charts

Pie charts



Definition

A *pie chart* is a circle divided into pieces proportional to the relative frequencies of the qualitative data.

► The steps to construct a pie-chart³

Step 1 Obtain a relative-frequency distribution of the data.

Step 2 Divide a circle into pieces proportional to the relative frequencies.

Step 3 Label the slices with the distinct values and their relative frequencies.

So, a pie chart, how do we construct a pie chart? Again, go back, we know how to construct a relative frequency table.

(Refer Slide Time: 02:19)

Statistics for Data Science -1
 └ Charts of categorical data
 └ Pie charts

Example



Use a protractor and the fact that there are 360° in a circle. Thus, for example, the first slice of the circle is obtained by marking off $0.4 \times 360 = 144^\circ$.

1. A,A,B,C,A,D,A,B,D,C

Category	Tally mark	Freq	Rel freq	Degrees
A		4	0.4	144
B		2	0.2	72
C		2	0.2	72
D		2	0.2	72
Total		10	1	360°

How do we obtain a pie chart? So you can see that I start by drawing a circle. Now, for example, the relative frequency of A is 0.4, I multiply that with the total number of degrees in a circle, which is 360. And I have a 144. This is 72. So, if I start from this, approximately this is somehow this is where I am going to have and this angle. I am not doing it exactly, but I know that this is going to be 144° .

Now the next thing I have is B which is at 72, C it is 72 this is 72° , this is 72° , and this is 72° . So, you can see what you can do is a way good way to have a pie chart is I can shade it with different colors, where color green represents the category D, color blue represents the category C, color purple represents the category B, and color orange represents the category A.

So, in a sense, what a pie chart gives us is, it gives us the share of a pie. In other sense, you can say that 40% of my data, which is the share of this pie is category A, the rest of them you can see the purple, the blue, and the green are almost same. They are same, in fact, they are same share, which is 20% each and that gives a share of a pie. So, whenever I want to actually show to my audience or to I want to display the share of a particular category, then I a pie chart is the most appropriate graphical display.

(Refer Slide Time: 05:11)

Statistics for Data Science -1
└ Charts of categorical data
 └ Pie charts

Pie chart in a google sheet

Step 1 Select/Highlight the cells having data you want to visualize.

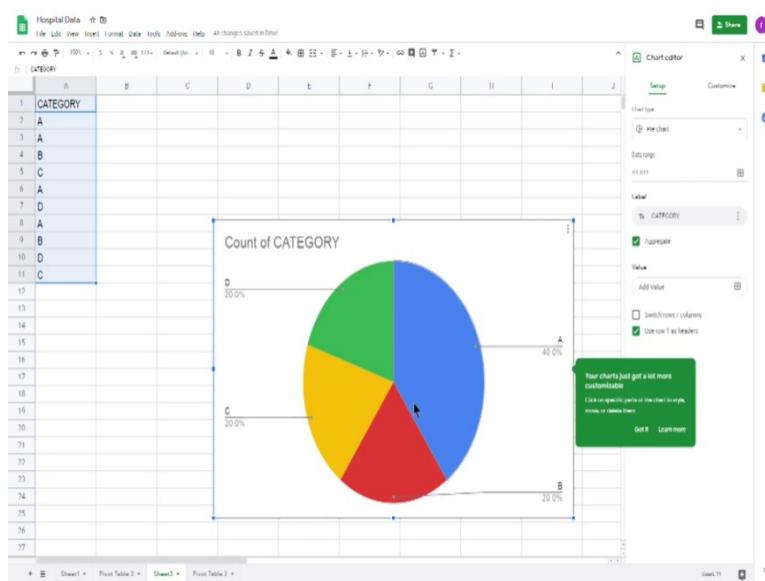
Step 2 Click the Insert Chart option in Google Sheets toolbar.

Step 3 Change the visualization type in Chart editor.

Step 4 Select in Chart Editor Chart type to Pie chart.

So you can see that for this, I can do a pie chart in Google Sheet as well.

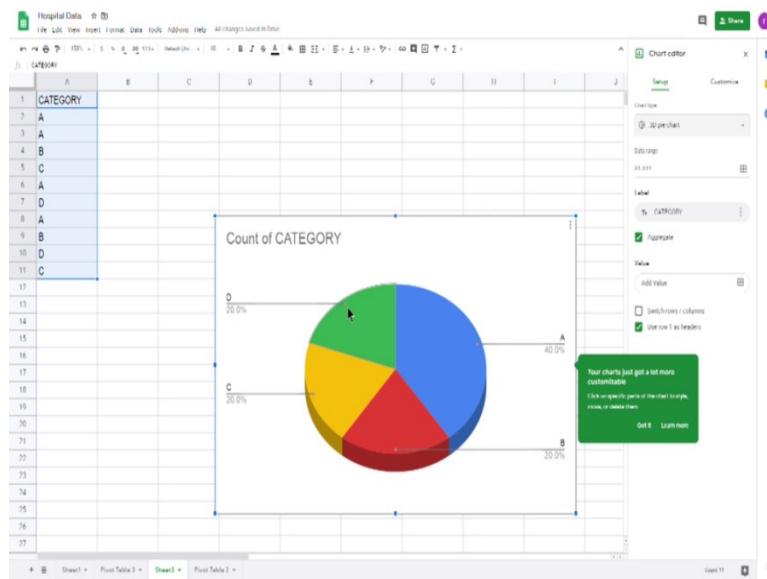
(Refer Slide Time: 05:16)



So, let us go and do a pie chart for the same data. So I have this as my data, I have this as my relative frequency. So, I go back to my data, which I have calculated here. So, you can see that, how do I do a pie chart again, I highlight the data I want to visualize, what is the data I want to visualize, I want to visualize this data. This is the hypothetical data I have created. I go to Insert Chart option.

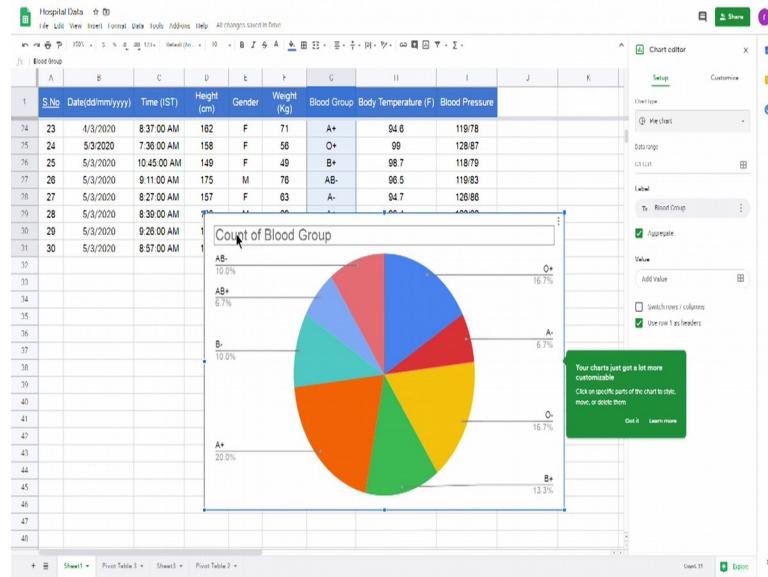
So, you have an insert option here. I have a chart option here. And you can see that this is precisely what we had earlier. A is 40%, B is 20%, C is 20%, and D is 20%. So, you can see that this is how my pie chart, you have the chart editor, within the chart editor, you can actually tell what is the kind of pie chart you want.

(Refer Slide Time: 06:16)



If you want a three dimensional pie chart, you can also click on that three dimensional. You have A again is 40%, B is 20%, C is 20%, and D is 20%. So, in Google Sheet, you change the visualization and chart editor to a pie chart and you get the pie chart. Now let us do the pie-chart for our blood group data.

(Refer Slide Time: 06:44)



So, this is my blood group data. Again, what I do, I have to select the datasets, I want to get hold of the distribution of my blood group. So again, I go to Insert, I go to a chart and within the chart, I am going to look at a pie chart, and you can see that this is the pie chart, which says O+ is 16.7%, A- is 6.7%, O- is again a 16.7%, A+ is 20%, B+ is 13.3, B- is a 10%, AB+ is 6.7%, and AB- is 10%.

So, you can see that how we can construct and what are we constructing, the variable is nothing but the count of the blood group. This is giving me a pie chart, which is giving me a distribution of the blood group. Of course, within this you can customize it, you can again, do what is the legend and all of that and these things would be taken up in the tutorial sessions on the discussion board.

(Refer Slide Time: 07:56)



1. A pie chart is used to show the proportions of a categorical variable.
2. A pie chart is a good way to show that one category makes up more than half of the total.



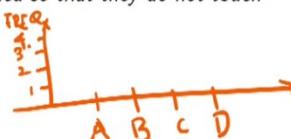
So by this time, you should have learned how to plot a pie chart. Remember, whenever the message is to show proportions of a particular categorical variable, a pie chart is a good way as it makes up too. And it will always tell you that the message, if one category, like in our example, where we had A made up for more than actually 144° is made up, you can see that out of, it has more than one way you can always use a pie chart to tell the share of a pie.

(Refer Slide Time: 08:34)



Definition

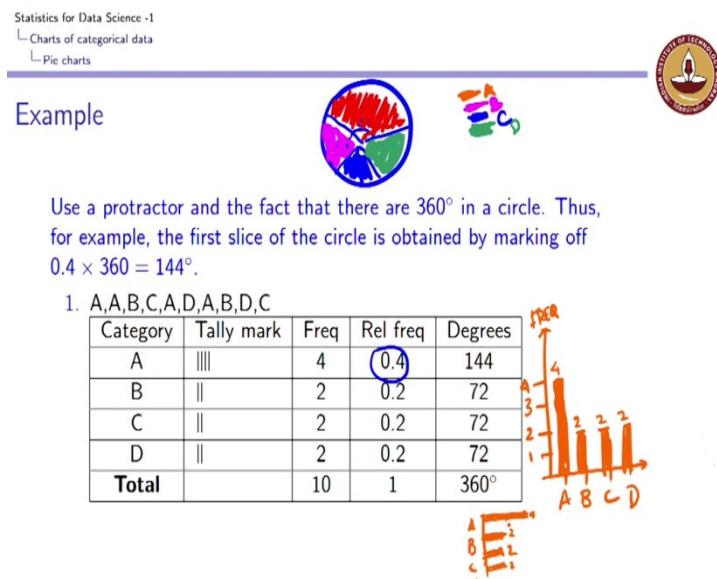
A bar chart displays the distinct values of the qualitative data on a horizontal axis and the relative frequencies (or frequencies or percents) of those values on a vertical axis. The frequency/relative frequency of each distinct value is represented by a vertical bar whose height is equal to the frequency/relative frequency of that value. The bars should be positioned so that they do not touch each other.



The next graphical summary which is also very important way, we said that when it comes to categorical data, the two most popular graphical displays are the pie charts and the bar chart. What is a bar chart? Again a bar chart, again, it displays distinct values of qualitative data on horizontal axis. So what do I have if I have a graph on my axis, I give the distinct value. For example, the distinct values were A, B, C, and D. These are the distinct values.

Now this, I really do not care whether it is a B, A, C, D because there is no order in this particular variable. But however, I need to be very clear as to what is the variable and the distinct values are given on the horizontal axis. On the vertical axis, I either can plot the frequencies or the relative frequency depending on what is my interest. We start with just a frequency. For example, if this were a 1, this is a 2, this is a 3, this is a 4.

(Refer Slide Time: 10:14)



Let us go back to our table, draw horizontal axis. Now, if I go back to my example I had here on the horizontal axis, I am just going to plot A, B, C and D. This is what I plot on the horizontal axis. On the vertical axis, I have 1, 2, 3, 4. So, A I draw a bar corresponding to A, the bar should be of equal width for each category, B is a 2, C is a 2, D is a 2. So here, I have the frequency, which is on my Y axis, I could have different colors, but I can have the same color also. But this is what is a typical bar chart.

(Refer Slide Time: 11:12)



Steps to construct a bar chart

To Construct a Bar Chart⁴

- Step 1 Obtain a frequency/relative-frequency distribution of the data.
- Step 2 Draw a horizontal axis on which to place the bars and a vertical axis on which to display the frequencies/relative frequencies.
- Step 3 For each distinct value, construct a vertical bar whose height equals the frequency/relative frequency of that value.
- Step 4 Label the bars with the distinct values, the horizontal axis with the name of the variable, and the vertical axis with "Frequency" / "Relative frequency."

⁴Weiss, Neil A. Introductory Statistics: Pearson New International Edition.
Pearson Education Limited, 2014.

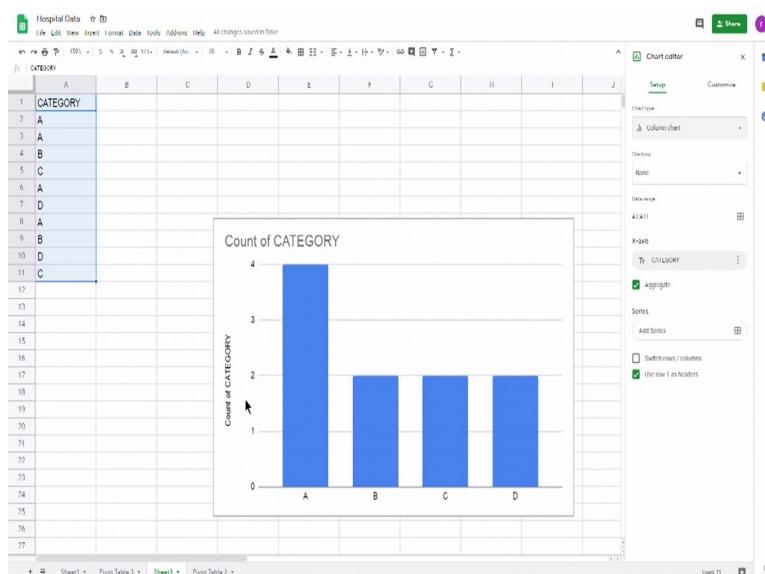
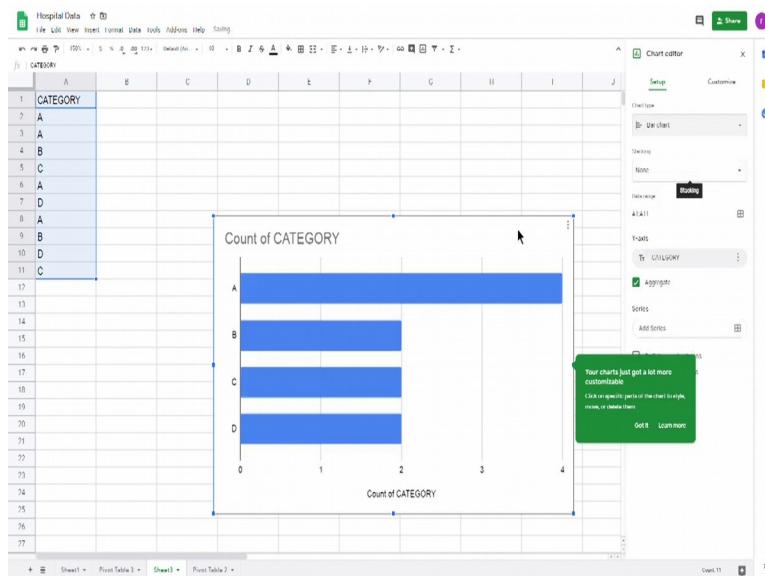


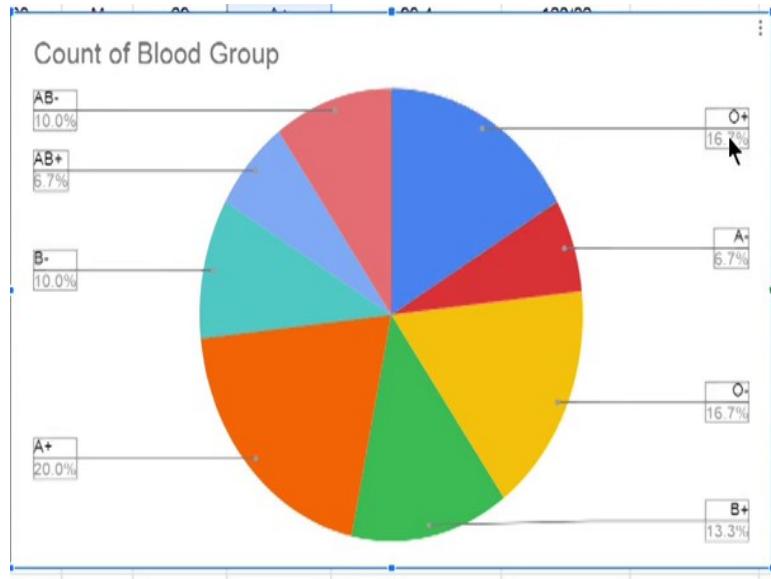
So what is needed, on the horizontal you have the bars, on the vertical the frequency. Label the bars so I can also go and one way I can annotate these bars, so I here I can right since I know the counts, just as I wrote the angles and the percentages for a pie chart, I can write a 4 here, or 2 here, or 2 here and a 2 here, which actually gives me the count of the Category A in this bar. So, this is a typical bar chart.

Now, a bar chart need not be only vertical, you could also have a horizontal bar chart where I have the categories on my axis which is A, B, C, D, and the counts C and D, and the counts which are given, here, I can have a A which is taking the value 4, so I have an axis which will give my value 4, this is B which is 2, C which is 2, and D which is 2. So, bar charts could be either vertical or horizontal. And depending on how you want to display your data, what is the message you want to give with your data, you can choose to have a vertical bar chart or a horizontal bar chart.

So, how do we create a bar chart in a Google Sheet, we go back to the data we have been working on. So I have, I highlight my data, that is my first step, which is given here.

(Refer Slide Time: 13:17)

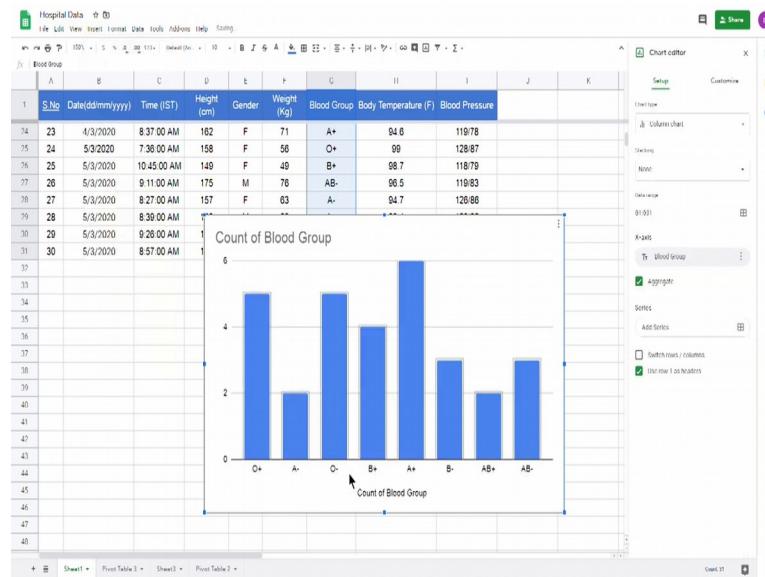




Once I highlight my data, I go to the Insert option with the chart, I choose a column chart here. Now you can see that I can choose either a count of category, which is a chart I said before, or I can also choose what is called a column chart. This is a column chart, just. So you can see that this is a column chart where again you can see on the y axis I have a count, I had 4 of Category A, I had 2 of Category B, I had 2 of category C, and 2 of category D.

Now, let us repeat that with our blood group data. If you go to the blood group data, again, I highlight the blood group data which is already highlighted. I go to insert. I go to chart.

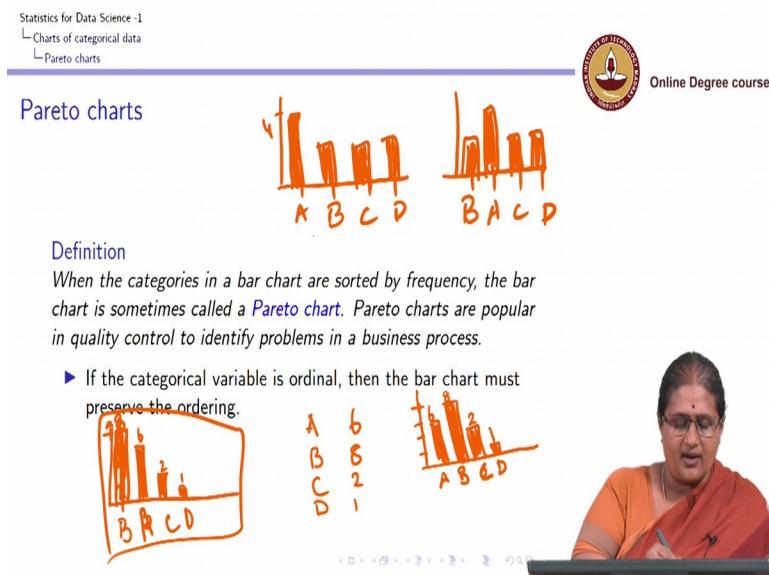
(Refer Slide Time: 14:02)



And you can see that this is exactly the count I had. And this is how this comes again from a relative frequency and my pie chart, you can see that the difference between the pie chart and the column chart is here. I know what is my count and I can annotate each one of the bars with exactly what is the count of people who are having that particular blood group. For example, there are 6 people out of 30 who have blood group A+, there are 2 people who have blood group A-, there are 2 people who have AB+ so you can easily see the count of people, there are 4 people who have B+.

What the pie chart gave us was the share of total, what is the relative frequency and you can see a bar chart is actually giving you the frequencies. You can also plot the relative frequency in a similar way.

(Refer Slide Time: 15:57)



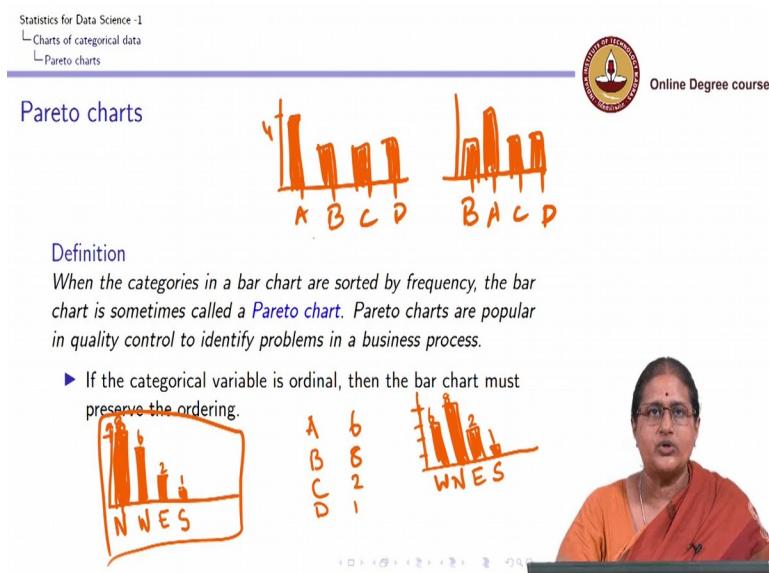
Now, many a time, what you might want to know is, for example, in this suppose I have been given this as a chart, I might want to have the bars which are arranged in a particular order. For example, here I see, A bar has the highest bar or the longest bar if it is a horizontal bar chart, next is both O- and O+, then I have B, B+ then I have AB-, B- and then I have A. So, I might want the bar chart to be arranged in a, as I already said, there is no problem in having a different order.

So in a sense, what I mean is when I have a bar chart, whether A, B, C, D. I have a 4, I have a 2, I have a 2, I have a 2. I could also displays display this as B, A, C, D, where this is a 2, this is a 4, this is a 2, and this is a 2. No problem with either of these displays because they convey the

same information and there is no order between A, B, C, and D. But however, in this I have the highest frequency appearing first and then I have the lower frequencies which are appearing.

For example, if I had a frequency distribution where A was 6, B was 8, C was 2, and D was 1. If I plot a bar chart with 2, 4, 6, 8, my A was a 6, category A was 6, category B was 8, category C is a 2, and category D is a 1. So, I have a 6, 8, 2, 1. This is giving me a count. A Pareto chart is something where I can just look at the category B which has the highest frequency, which is a 8, then comes B, A which is 6, then come C which is 2, and last is D which is 1.

(Refer Slide Time: 18:53)



Now when would a chart of this kind be of use to us. Suppose instead of A, B, C, D I had our states, suppose or I had our regions, suppose I had the northern region, the southern region, the eastern region and the western region. Broadly classify any particular geographical area into four particular regions, and I am interested in knowing what is the employment rate or what is the number of students who are passed in each region.

If I present a chart in this way, and for just hypothetically this represents 8000 students, 6000 students, 2000 students, and 1000 students and I have a distribution. Again, just for hypothetical purposes, this is the northern, this is the western, this is the eastern, this is the southern region. So at a snapshot, I can immediately see, now this data would have been the same as this data, where this was the southern region, this was the eastern region, this was the northern region and this was the western region.

Both the data convey the same message, whereas the Pareto chart in some sense, gives me the idea at a snapshot. I can know that the northern region has the highest number of people followed by the western, then the Eastern and the southern. So, you can see that the bars are in descending order. I can also have the bars in ascending order. Again, it depends on what is the message you want to convey. Hence, in this chart, these type of charts are referred to as Pareto charts.

(Refer Slide Time: 20:03)

Pareto charts

Definition

When the categories in a bar chart are sorted by frequency, the bar chart is sometimes called a **Pareto chart**. Pareto charts are popular in quality control to identify problems in a business process.

- ▶ If the categorical variable is ordinal, then the bar chart must preserve the ordering.
 - **NOMINAL → ORDINAL → ORDER**
 - $\rightarrow S, M, L$

Online Degree course

Now recall when we talked about categorical variable, we said categorical variable can be measured in two scales and that scales I call them as a nominal scale and I also call them as an ordinal scale. Now the key difference between a nominal scale and an ordinal scale is in this I just have name or values, whereas ordinal Scale I have a natural order. For example, if I am looking at small, medium, large, there is a natural order even though these are categorical and I don't have numerical values assigned to them.

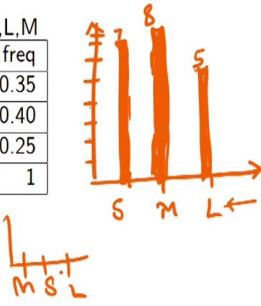
(Refer Slide Time: 20:47)



Example- ordinal data

The T-shirt sizes (Small-S, Medium-M, Large-L) of twenty students is listed below:

Size	Tally mark	Freq	Relative freq
Small		7	0.35
Medium		8	0.40
Large		5	0.25
Total		20	1



Now when the categorical variable is ordinal, like I have in this data, I have T-shirt sizes small, medium, large of 20 students. So, I have constructed the tally mark. So, I have 7 students who have small T-shirts, medium are 8, large are 5. So, I can construct my frequency or my bar chart for this data. So I have the size of the T shirt, which is small, medium, and large, the sizes, so I have a 1,2,3,4,5,6,7,8. I have a size 7 for my small, a size 8 for my medium.

The bar should be of the equal width since it is free hand. The bars do not appear to be but you can see that your Google Sheets always provide bars of the same width. 3, 4, 1, 2, 3, 4, 5, and large is 5. I further annotate them, this is a 7, this is a 8, and this is a 5. So, whenever there is an ordinal it is good to maintain the order of the categorical data.

For example, I do not want to have a chart which is medium, small, and large, because the order of this categorical data is not maintained. So, even within a bar chart for a categorical data, if you have an order, please maintain that order of the categorical data.

(Refer Slide Time: 22:35)

Summary



1. A bar chart is used to show the frequencies/relative frequencies of a categorical variable.
2. If ordinal, the order of categories is preserved.
3. The bars can be oriented either horizontally or vertically.
4. A Pareto chart is a bar chart where the categories are sorted by frequency.

So in summary, we know, a bar chart can be used for frequencies or relative frequency depending on what you want to convey. On the x axis, I have the categories, and on the y axis I have the frequencies which are the counts if I am plotting the count, or the relative frequency if I am looking at relative frequency, I can either have a horizontal bar chart, where I have, a horizontal bar chart would look something of this kind or a vertical bar chart, which would look something of this kind. A Pareto chart is where the categories are sorted. And if you have ordinal data please try and preserve the order of the categories.

Statistics for Data Science - 1
Professor. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture No. 2.3
Describing Categorical Data - Best Practices While Graphing Data - 1

So, what we are going to do now is to understand certain best practices about drawing these pie charts and bar charts what we have learned so far.

(Refer Slide Time: 00:24)



The screenshot shows a presentation slide with a navigation menu on the left. The menu items are: Statistics for Data Science - 1, Best Practices and Misleading graphs, and Best practices. Below the menu, the text "Know your purpose" is displayed. To the right of the text is a circular logo of the Indian Institute of Technology, Madras, featuring a lamp and the university's name.

- ▶ Have a purpose for every table or graph you create
 - ▶ Choose the table/graph to serve the purpose.
- ▶ Pie charts are best to use when you are trying to compare parts of a whole.
- ▶ Bar graphs are used to compare things between different groups.



Now even before we actually go and understand about what is required from a bar chart and pie chart or a frequency table, the first thing we need to understand is what is the purpose? Now when I said purpose it does not mean or it does not necessarily mean that every dataset should convey a message. What is the purpose? We asked our question what is the purpose? For example if I just have a data say I have a data which is A, B, C, A, B, D, E, A or D, A, B, B this is just a categorical data.

Now if I have this data, now the questions I need to ask is what is the purpose of this data? Suppose this data is just a set of states and nothing more than that suppose this is Andhra Pradesh, this is Bihar, this is Chhattisgarh and all of that and I am looking at I am collecting data or I am asking every person who is entering a particular room which state they belong to and this is the data I have.

Now once I collect this data I need to understand what are the questions I want to ask from the data? The first thing is if I want to just tabulate this data. So, I look at it, I prepare a table like the way we discussed in the last time I say people from state from A, people from state B, people from state D, people from state D this is the category I am looking at and then after I have the count which I called it is also the frequency.

Once I have this count this could be say 1, 2, 3, I have 3 I have 1, 2, 3, 4, 4. C is just 1 and D is 2. So this is the way I construct what is called a frequency table. So, now the question is what is the purpose from this if my purpose is just to count and represent it as a table I go in for a frequency table. However, if my purpose is to come up with a tabulation I want to compare how each state does with the other.

Then I might want to go for a bar chart because bar chart helps me in comparison. However, if my purpose is to know what is the share of each state then I will go in for relative frequencies which is depicted by a pie chart. So, what we are going to do today is to first understand when to use a pie chart. You use a pie chart when we are trying to compare parts of a whole.

So, if I have class of say 100 students and I want to know what is the regional distribution, then after a pie chart is the right chart for me to use to compare this. However, if I want to compare things I want to know what is the exact count of how many are from each region then a bar chart is more appropriate. So, the first thing when would you use a table. We would use a table when we are interested.

Suppose I have a lot of categories then what would happen is both the bar chart and pie chart would appear to be entirely cluttered. If I want to represent the entire data then perhaps a table is more appropriate. So, even before we go to summarize data the first question we need to ask is what is the purpose of my summary, analysis, what is it I want to convey. Choose the table or graph to serve this purpose.

(Refer Slide Time: 04:20)

Statistics for Data Science -1
└ Best Practices and Misleading graphs
 └ Best practices



Label your data

- ▶ Label your chart to show the categories and indicate whether some have been combined or omitted.
- ▶ Name the bars in a bar chart.
- ▶ Name the slices in a pie chart.
- ▶ If you have omitted some of the cases, make sure the label of the plot defines the collection that is summarized.



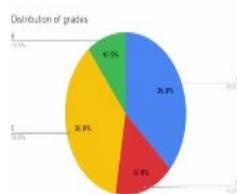
The next very important thing is what we saw in the earlier class was how to create both the bar chart and the pie chart. Now we are going to spend some time to understand how to label the bar charts or how to label the charts.

(Refer Slide Time: 04:41)

Statistics for Data Science -1
└ Best Practices and Misleading graphs
 └ Best practices

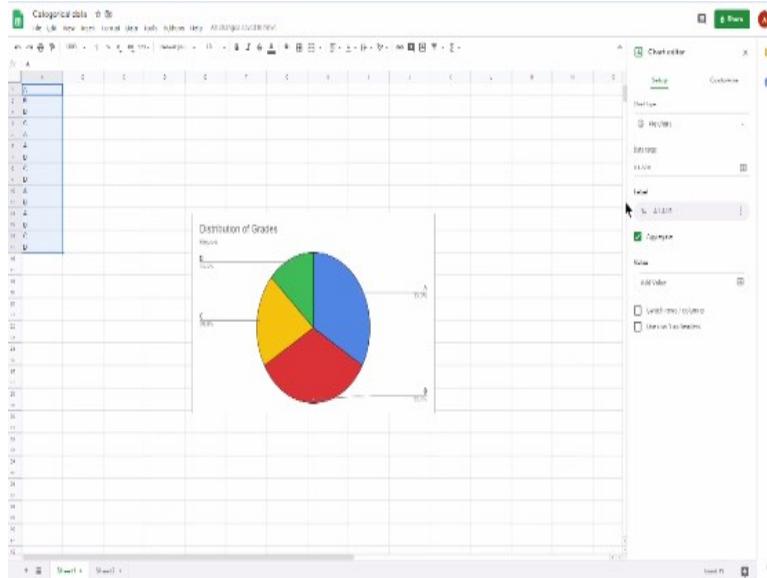


Label your data



So, even before we label the chart.

(Refer Slide Time: 04:43)



So, now let me just go back to our Google sheet and type in a particular data. Suppose, this is the data I have just typed in I just want to create a chart of this data, so I go to insert, I go to chart and you can see that I have a pie chart which is inserted here. Now within the pie chart you can see that I have a tab which says it is customized and that is on your right hand corner. Now we can see that, here this pie chart I have what are the pie slices.

I know I can put up a title this title is something currently the title is count I can just put a title which says distribution of some grades. If A, B, C, D represents grades I can just say distribution of grades. Now typically if you want a subtitle you can add a subtitle. For now I am just adding a subtitle which is region just for the sake of it and you can see the distribution of grades region appears here.

Now it always helps us to then look at what are the legends you can want you can put the legend wherever you want and I am sure that we will have a separate tutorial to actually tell how different legends would look later, but at this point of time what I want you to tell is you can see what is the color whether it is auto and you can keep adding the colors and the text colors or you can just put it auto the title font size.

You can look at what are the chart this one and suppose I put 25 you can see that you are actually moving a particular slice out of the particular thing. One point I want to make here what I want to clarify is when you are clicking two colors is not advisable because when you click two colors

it is really you do not have any change between these two colors. So it is always advisable to retain one particular color.

Now suppose I do the same thing for each of these pie slices I go and I do each of them. This even though it is possible you see that this does not make any sense. The reason why a chart like this does not make any sense it is not actually the region the purpose of a pie chart is to help you understand what is the share of a pie. Now when just for decorative purposes if people start demonstrating it this way you can see that this does not actually convey what you want to say.

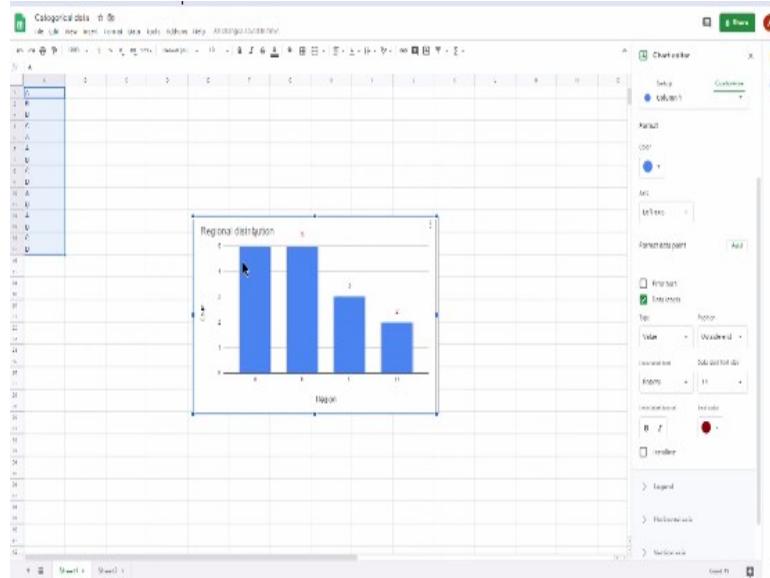
So I would strongly advise against a visualization where you are actually moving every pie slice rather I would prefer that unless essential do not have to change the distance from the center and have everything maintained so far we have a nice pie chart. Again this would tell us how to label it whether you want it in the label what is the font size and whether you want it as a in italics or whether you want it bold you can see as I keep changing here.

So, all of that you can do it later. Now another thing is if you want borders you can add a border or adding a border would always help us. You can always decide whether you want a border or not. There is something else in the layout which is called a 3D. Again, I do not think that 3D is giving us any extra value, but you can always have a pie chart as a three dimensional chart also.

You use it whenever you want it, but in my opinion a simple pie chart conveys more than what is essential. So, this is about how you are going to come up with chart title, chart subtitle and what are the default size for each pie slice you can see what is the distance from the center. You can customize it and then after you can add values whenever you want, you can switch rows and columns here but here since I have only one column I do not have to do that.

Now the same thing, the same data I can choose here for a same data I can choose the chart which I call a bar chart or a column chart here.

(Refer Slide Time: 10:12)



Now again in a column chart I can customize the column chart. First thing I add the subtitle and title, I can add the chart title again I am going to add the chart title as regional distribution assuming these are A, B, C, Ds are regions. I can add a subtitle if I want to. Here I need to add a horizontal axis title. In the horizontal axis title I can just add region because my A, B, C, D are actually representing regions.

Again within a region I can tell what is the font I needed if I need a very big font size and if I need it in bold I can do all those things here. I also have a vertical axis, in the vertical axis I just add a count. The vertical axis this count is given. So, this count is given 1, 2, 3, 4, 5 so another thing which I would like to add is I can add data labels where data labels it tells me actually what is the number of counts or observations for each category.

Now this auto labels can be in any size, too big or it can be 5. The text color I can choose the text color I need for my data labels and I also can tell what is the position. I always prefer it at the outside end or I can have it at the inside base wherever you want. This I am not going to have there is no particular actual to be done it or this is what you can have, you can see that. You have the region with the distribution which is at the outer end.

(Refer Slide Time: 11:55)

Statistics for Data Science -2
└ Best Practices and Misleading graphs
 └ Best practices

Label your data

Distribution of grades

Grade Category	Percentage (%)
A	36.8%
B	15.8%
C	36.8%
D	10.5%

Distribution of grades

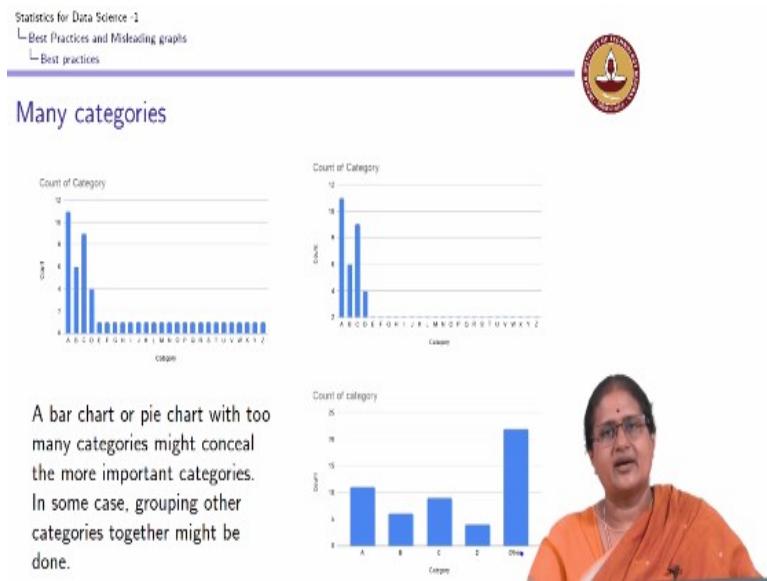
Grade category	Count
A	7
B	3
C	7
D	2

So, what the first thing which we need to understand is to be it a pie chart or a bar chart. First thing is to label or annotate your data because only when we label or annotate the data there is a better visualization or it communicates the idea better. So, here again if these were A, B, C, D I have label my chart distribution of grades and this gives me the distribution of grades which is 36.8 % of A 15.8 % of B, 36.8% of C and 10.5% of D.

Similarly, this is giving me the count. I have 7 people who have got A, 7 people who have got C whereas 3 people who have got B and 2 who have got D. So, you can see that this is telling the share of a particular grade whereas this is giving a count. So, whenever you want to represent data in either a pie chart or a bar chart the first thing you do is label or annotate your chart.

Now what you see the horizontal lines here these are called gridlines. Now you can also choose whether to have the grid lines or not that you can go and you can choose again in your data whether you would want to have gridlines or not and that can be done here whether you want the grid lines or not you can actually choose it here. So, when I have the option of no gridlines I get a chart without gridlines. Now suppose I am giving you another data set.

(Refer Slide Time: 13:53)



Now if you look at this data set it tells me that I have about 11 of category A, 6 of category B, about 9 of category C and about 4 of category D and I have many, many, many, many more categories each one of them they have only one of each such category. Now where could such data come up from? Now suppose for example I am asking some hundred people who is your favorite cricketer?

There would be an overwhelming response that Sachin Tendulkar is their favorite cricketer, some of them might choose Virat Kohli others might choose say Zaheer Khan some might choose MS Dhoni, but then afterwards you have all these little, little guys only two of them might choose say Rahul Dravid, two might choose KL Rahul, two might choose Ashwin and all of this.

So, this is in a sense that the major choice could be the first 4 cricketers and others it is not that my entire 100 people are going to only choose among these 4 cricketers, but I have a distribution. So, in this case what you see is when I immediately look at a graph of this kind I find too cluttered because there are too many categories. Now this could also be a case where I am looking at distribution of industries in particular states.

There could be only 4 states which sees maximum number of industries and I have a splatter of industries in other states. So, now we have too many categories and the bar chart actually looks

very cluttered and it is not conveying what I am looking for it to convey. What do we do in these cases? So, a bar chart or pie chart with too many categories might actually conceal the more important categories.

So, one way to do it is do not ignore the category. So, the chart now currently what chart I have not plotted those categories I have only given a command that plot the bars only if it has more than 2 count. So, you can see that all my categories have been which had lesser than 2 counts are not figured in the bar chart at all. What I would suggest is go to a category here where the others are clubbed into a major category.

Now if you do this it conveys two important things. One is you are not excluding any data and the second important thing which it conveys is that even though this bar chart says 11 of the total come from category A and says that 9 from category B what this chart conveys is of the total number you have a significant number that comes from the smaller categories and that you can see is about 22.

So, this gives so eliminating or not giving this information actually does not give you the entire story. So, this is one way where you can club in all the smaller categories or categories which have very little count to call it as an other category and this when you portray many categories as an other category it gives you the entire story.

Statistics for Data Science - 1
Professor. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture No. 2.4
Describing Categorical Data - Best Practices While Graphing Data - 2

(Refer Slide Time: 00:14)

Statistics for Data Science - 1
└ Best Practices and Misleading graphs
 └ Best practices

The area principle



- ▶ Displays of data must obey a fundamental rule called the area principle⁵.
- ▶ The **area principle** says that the area occupied by a part of the graph should correspond to the amount of data it represents.
- ▶ Violations of the area principle are a common way to mislead with statistics.

⁵Stine, Robert, and Dean Foster. Statistics for Business: Decision Making and. Addison-Wesley. 2011.



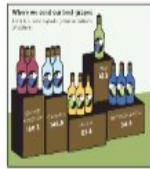
Now whenever we come to display a data there is a fundamental rule called the area principle. Now, what is this area principle actually tell us. When we look at the area principle the area principle states that the area occupied by a part of the graph should correspond to the amount of data it represents. What do we mean by this?

(Refer Slide Time: 00:42)

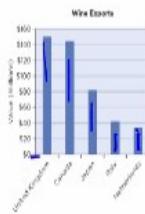


Misleading graphs: violating area principle

- Decorated graphics: Charts decorated to attract attention often violate the area principle⁶



- No baseline and the chart shows bottles on top of labeled boxes of various sizes and shapes.



- Obey's area principle and accurate



⁶Stine, Robert, and Dean Foster. Statistics for Business: Decision Making

You can look at the first thing is when I have charts which are decorated to attract attention. For example let us look at this particular chart now what this chart gives us is the total wine exports in the United States value in millions of dollars. Now this is more like a infographic, but what does this tell? It tells us United Kingdom is about 150.3, Canada is 146.8, Japan is 82.8, Italy is 42.5 and Netherlands is 34.4.

You can immediately see that there is no baseline, it show bottles on top of label boxes. Now even the boxes are not of any uniform shape they are of various shapes and sizes. So, when you look at this chart it does not convey anything which is actually I can make a meaning out of. Whatever I need to convey is just a total US wine exports to each one of these countries and I can construct this using a bar chart where my categories are again United Kingdom, Canada, Japan, Italy and Netherlands.

I have labeled each one of my categories. I can see that United Kingdom has a value in 150.3 million dollars, Canada is about 146.8 million dollars, 82.8 to Japan, 42.5 to Italy and 34.4 to Netherlands. Now the chart on the right hand side obeys what we call the area principle. It is accurate, it has a baseline this is the baseline of this chart. So, you can see that this is the baseline of this chart.

It has an everything is it is actually consistent the width of the bars for each countries is equal and I can have a vertical scale and on the vertical scale I have the value which is given and hence

I obeys what I refer to as the area principle where the area occupied by the graph and what is the area occupied by the graph it is this area, it is this area, it is this area, it is this area and this area it is proportional to the data that is being presented.

(Refer Slide Time: 03:21)

► Another common violation is when the baseline of a bar chart is not at zero.

Regional distribution

Region	Count
South	100
North	75
West	25
East	50

Regional distribution

Region	Count
South	100
North	75
West	25
East	50

Left graph exaggerates the number coming from the South and North. Graph on right shows same data with the baseline at zero.

INTEGRITY

The next way people mislead with graph is through use of what we call as truncated graph. Now what is a truncated graph? Now let me show one thing is where the baseline of a bar chart is not at 0. What do we mean by baseline of a bar chart is not at 0? Now let me show you this graph now again this is a regional wise distribution. I have four regions South, West, North and East.

Now when I portrayed this data and this graph to you, you can immediately see there are 500 people and I have just the data is which region do the 500 people come from, that is the data. Now when I portray or I show this graph the immediate thing or the immediate response or a person who looks at this data without looking at the access is to imagine that these people from South and North are much higher than the people from West or East.

That is what this message this graph conveys, but when we look more carefully at the data you see that this is the baseline and it starts at 0 whereas here it starts at 75. Now these two graphs are actually the same data. What do I mean by same data? 500 students each one of them tell which state, region they come from whether it is a South, whether it is a West, whether it is a North or the East.

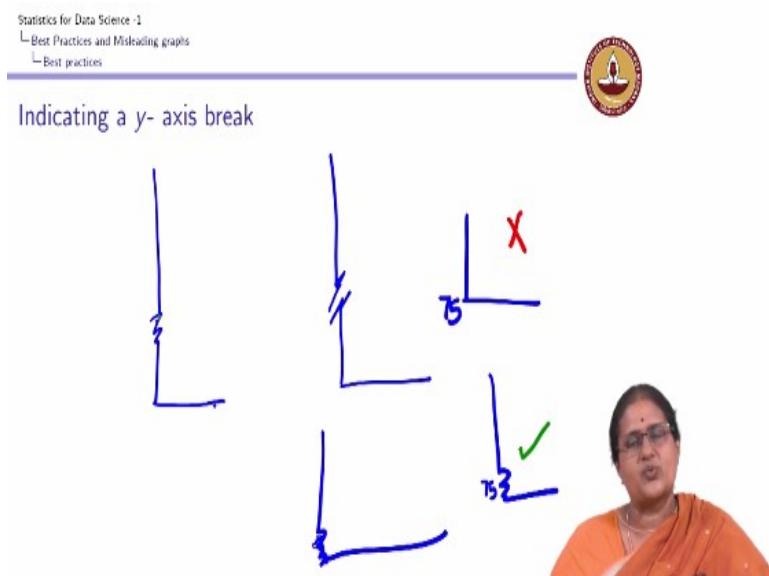
When you look at this graph, you feel that the South and North are the ones from where majority of the student come. Even though majority come it seems that the North and East are negligible

to the distribution from North and South, whereas this graph conveys a different story. So, the left graph or this graph exaggerate the number whereas this graph shows the same baseline at 0.

Now, this is what is recommended because this is where data integrity is maintained and the actual story is said. Now where does this matter, now when you are showing a growth and everything and you truncate the graph or you miss the present data or you mislead, people will actually attribute a wrong story to this even though both of them represents the same data. Visually you see that this tells a different story than what do the one on the right says.

So, the second thing is whenever you truncate graph there is a loss in information and it is a violation this has to be avoided.

(Refer Slide Time: 06:42)



Now some textbooks and some people say that whenever you have a truncated graph or something you introduce a y axis break. So, if you are starting and you are breaking the axis you can introduce it either by doing this where I am telling that I am introducing a break in my y axis or you can show that this is I am not starting from 0 there is a break and I am starting from a higher value.

For example, if I wanted to start from 75 I should have indicated that I am starting from 75 instead of shifting the graph to 75. So, this is incorrect whereas this is this is the right way to go

about it. So, whenever you are altering or manipulating with the y axis, indicate to the reader that you have manipulated or you have introduced a y axis break.

Statistics for Data Science - 1
Professor. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture No. 2.5
Mode and Median

(Refer Slide Time: 0:14)

Category	REL FREQ
A	22.5 → 23
B	35.5 → 36
C	12.5 → 13
D	11.1 → 11
E	18.5 → 19
	100 102

⁷Stine, Robert, and Dean Foster. Statistics for Business: Decision Making and. Addison-Wesley, 2011.

Now the other thing which is very important to check is whether you are introducing round off errors. Now what do I mean by round off errors? Now suppose I have a data set where I am having 22.73 I had 35.74, 11.30 and I have 18.23. Now suppose these are the relative frequencies of 5 categories I am naming the Category A, B, C, D, and E and these are the relative frequencies.

Many a time we are tempted to round them off. So, I just do a 30, 20, 35 dwell, I can look at 11.3 and it is 11 point 11.20, 11.53 and 19. A very quickly we look at this is suppose I have 12.5 and 18.03, 12.5. And I happen to round it off, I round it at 13 and things like that. So, you can see that when you round this data, so I am round in 22.5 to 23, 35.5 to 36, 12.5 to 13, 11 remains at 11 and 18.5 to 19.

What happens in this case is you can quickly see that now, in this case what happened in the earlier cases I had the total 22, 35, 12. So, when I round it up, I get a 23, I get a 36, I get a 13, I

get 11 and I get a 19. And you can see that what you could, what you achieve by rounding up is my total here is 100% whereas my total here is 102.

(Refer Slide Time: 2:36)

Statistics for Data Science -1
└ Best Practices and Misleading graphs
 └ Best practices

Round-off errors

- ▶ Important to check for round-off errors.
- ▶ When table entries are percentages or proportions, the total may sum to a value slightly different from 100% or 1. This might result in a pie chart where the total does not add up⁷.

⁷Stine, Robert, and Dean Foster. Statistics for Business: Decision Making and. Addison-Wesley, 2011.

Categorical data

22.5	23	
35.5	36	
12.5	13	
11	11	
18.5	19	
100	102	

What is the implication of this, when you have actually table entries that are percentages or proportions, as we saw in the earlier case, what was happening is for simple data of this kind, you can see that I actually what is happening in this case is for simple data of this kind by rounding it off, you can see that this total is actually becoming actually the total becomes

changes from 100 to 102. So what would happen to this? So, look at a case where the data was rounded up, you can see 20 plus 15 is 35, 35 plus again, a 25.

Here I have the 20 I have 15, I have 14, I have 11, I have a 10 and at 31, you can see that this does not add up to 100. And you can see that when it does not add up to 100. This is not forming what I call a pie chart. So, it should be extremely careful when you are actually rounding up. So, you have an error here, where my data is actually adding up to 101 and not to 100. So, what you should be very careful about is when you are actually rounding off you should be careful to see that the round off errors are avoided.

(Refer Slide Time: 4:12)

Statistics for Data Science -1
└ Best Practices and Misleading graphs
 └ Best practices

Sectional summary



1. Know your purpose and choose table/graph appropriately
2. Label your charts
3. Handle multiple categories appropriately.
4. Respect area principle
 - 4.1 Avoid overly decorated graphs
 - 4.2 Avoid truncated graphs- use special symbols to indicate vertical axis has been modified.
 - 4.3 Check for round-off errors .



So, the first thing which we have learned so far is whenever you want to graphically or summarize your data through a table or graph know the purpose would be you want to label your charts or annotate them, handle multiple categories appropriately, even if the count of a particular category is negligible, combine categories as together of all categories, which have a small count, respect area principle, avoid overly decorated graphs. Avoid truncated graphs, because truncated graphs are mostly misleading. And finally check for round off errors.

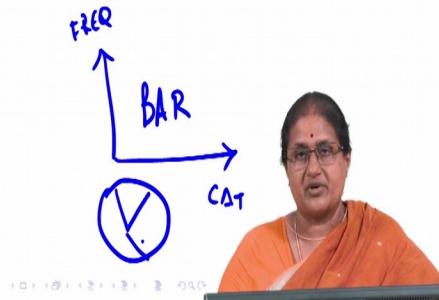
(Refer Slide Time: 5:03)

Statistics for Data Science -1
L – Mode and Median

Summarizing categorical data

TABLE	CAT	FREQ	REL FREQ
-------	-----	------	----------

- ▶ Graphical summaries of categorical data: bar chart and pie chart.



Now, the next thing which you are going to discuss is, So, far we have looked at graphical summaries and to tabulated data that is we looked at a frequency table, when we looked at a frequency table we talked about a frequency table where I have a category, I have a frequency, I have relate frequency. This is what I call was a frequency table.

Now, when I have when I plot category on my x axis with the frequency or count, I refer to it as a bar chart. And when I actually look at the relative frequency distribution, I call it a pie chart. This is what we have seen so far. And in the last section we saw how to label a pie chart how to label a bar chart and what are the certain things which we need to take into consideration when we are actually graphically summarizing data.

(Refer Slide Time: 6:04)

Statistics for Data Science -1
L – Mode and Median

Summarizing categorical data

SCALES

NOMINAL] CATEGORICAL

ORDINAL

INTERVAL] NUMERICAL

RATIO

- ▶ Graphical summaries of categorical data: bar chart and pie chart.
- ▶ Need for a compact measure.
- ▶ Numbers that are used to describe data sets are called descriptive measures.



However, we need a compact measure sometimes to describe our data, what do we mean by disc compact measures. So, numbers that are used to describe data sets are called descriptive measures. Now, you may recall that when we looked at different types of variable we said the types of variables were nominal ordinal interval and ratio. We label these as the scales of measurement, where these two are basically when the variable is categorical or qualitative in nature and this is when I have numerical or quantitative variable.

Since now we are focused on what we are looking at now is the nominal or the categorical variable and you are looking at how to summarize the categorical variable. And what we looked at here so far is we looked at the graphical summaries when we looked at graphical summaries, we are looking at bar chart and pie chart. Now the question is, do I have any descriptive measure to describe datasets where my variable is categorical in nature? The answer is yes.

Now again you go back to your nominal and ordinal variables, we said that we cannot have any arithmetic operations described on this except for counting the number of observations in a particular category. The difference between nominal and ordinal is there is an order in the categorical variable for example, sizes of a T shirt from small, medium large there is an order good, excellent superlative. There is an order good bad, ugly, there is an order if ABCD represents great A is better than B is better than C is better than D again the represents an order.

(Refer Slide Time: 8:32)



Summarizing categorical data

- ▶ Graphical summaries of categorical data: bar chart and pie chart.
- ▶ Need for a compact measure.
- ▶ Numbers that are used to describe data sets are called descriptive measures.
- ▶ Descriptive measures that indicate where the center or most typical value of a data set lies are called measures of central tendency

So, the question is, is there a number that I can use to describe these data sets? The answer is yes. When I describe the data set, the most typical value of a data set where the centre or most typical value of a data set lie is generally referred to as a measure of central tendency. That is where the centre or the typical value of a data set lies is called a measure of central tendency.

So, when we talk about a categorical variable, what is the measure of central tendency that you are referring to? I know that I cannot do any arithmetic operation on it.

(Refer Slide Time: 9:17)

Statistics for Data Science -1
L – Mode and Median

Mode

ABCDAABDCABC A

Definition
The mode of a categorical variable is the most common category, the category with the highest frequency.

The mode labels

- ▶ The longest bar in a bar chart
- ▶ The widest slice in a pie chart.
- ▶ In a Pareto chart, the mode is the first category shown.

CAT FREQ REL

CAT	FREQ	REL
A	6	0.4
B	5	0.33
C	4	0.27
D	2	0.13

Bar chart showing frequency of categories A, B, C, D.

Pareto chart showing relative frequency of categories A, B, C, D.

Speaker: [A woman in an orange sari speaking]

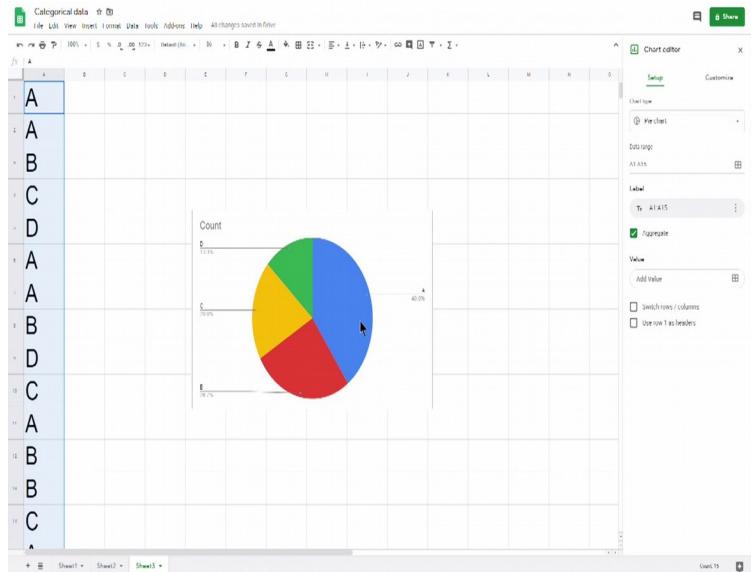
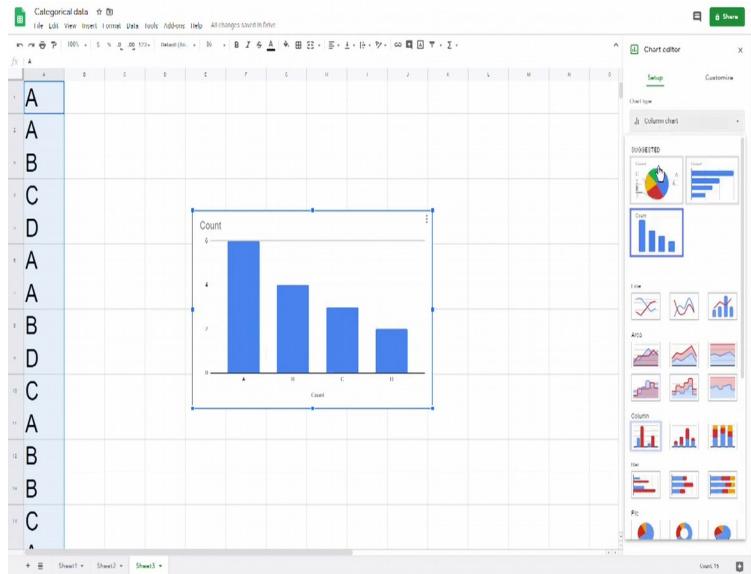
So, the first measure of central tendency is to count what is that category which has the highest frequency or highest count? So, if I have a data which is A, A, B, C, D, A, A, B, D, C, A, A, B, B, C, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 and I have a 15 again categories.

The category I have here is A, B, C, D, the frequency the way looked at is A11, B, C, D, A, A, B, D C, A, B, C, B. So, I can see that the frequency 5, 5, 3, 2 now instead of B here, if I had a A, this would have been a 6, this would have been a 4 I would have had 15 observations with this as my frequency table. Now, what is the mode of this data set? The mode of this data set is this value A because it is has the highest frequency.

So, you can see that the mode of this data set is this category A. Now, when we are plotting charts, again I know that I can plot a chart here, this is the Category A, B, C, D, A 1, 2, 3, 4, 5, 6 this is the count or the frequency A is 6, B is 4 6 4 C is 3, D is 2. So, this A is the mode because it has the highest frequency and this is the longest or the tallest bar in the bar chart.

Now, similarly, if I am to come up with a pie chart, again I look at the relative frequencies we have seen in the earlier classes how I find the relative frequency 6 by 15 four by 15, 15 and 2 by 15 I find out the angles and then you can also check and you can find out that the relative frequency.

(Refer Slide Time: 12:10)



So, if I am going to plot this data, so, I look at this data, so, it is a A, A, B, C, D. So, I go to this I have my data A, A, B, C, D, A, A, B, C, D, A, A, A, B, D, C. So, if you look at this data here, so, this is the data which we just had, I look at this data so, this is the data for this data insert a pie chart. You can see that A which is here is the largest pie, go to a bar chart. A again has a count of 6 B has a count of 4, C has a count of 3 and D has a count of 2, the length of the longest bar and the largest by both of them are the mode.

Now again, if you look at the bar chart for this case, here the A, B C, D are actually this is 6, B, C, D, this is a Pareto chart, but in case I had something of this kind, this is not a Pareto chart because the Category B is appearing first and then a then C and D. This is just a bar chart which

is actually listing these categories B, A, C, D, but when I have a Pareto chart, the Pareto chart for it actually the mode is the first category shown.

This is a Pareto chart the mode this is the first category that is shown and it is a Pareto chart. So, this is about what is a mode, the mode is the most common category with the highest frequency.

(Refer Slide Time: 14:09)

Statistics for Data Science - 1
L – Mode and Median

Example

◀ Let consider the example A,A,B,C,A,D,A,B,C,C, A,B,C,D,A
◀ The longest bar in a bar chart

Category	Count
A	6
B	3
C	4
D	2

The most common category is "A"

Navigation icons: back, forward, search, etc.

Now, this is the data which we are talking about, the most common category here is A something that you can notice in this data is the most common category is A, I have C, so, A, I have 6, I have B, which is 3, I have C which is 4, I have D which is 2 and A is my mode.

(Refer Slide Time: 14:36)

Statistics for Data Science -1
L—Mode and Median

Example

- ▶ Let consider the example A,A,B,C,A,D,A,B,C,C, A,B,C,D,A
- ▶ The widest slice in a pie chart.

The most common category is "A"

A woman in an orange sari is visible on the right side of the slide.

Similarly, for the same data, I have A which is 40% the most common category is A.

(Refer Slide Time: 14:44)

Statistics for Data Science -1
L—Mode and Median

Bimodal and multimodal data

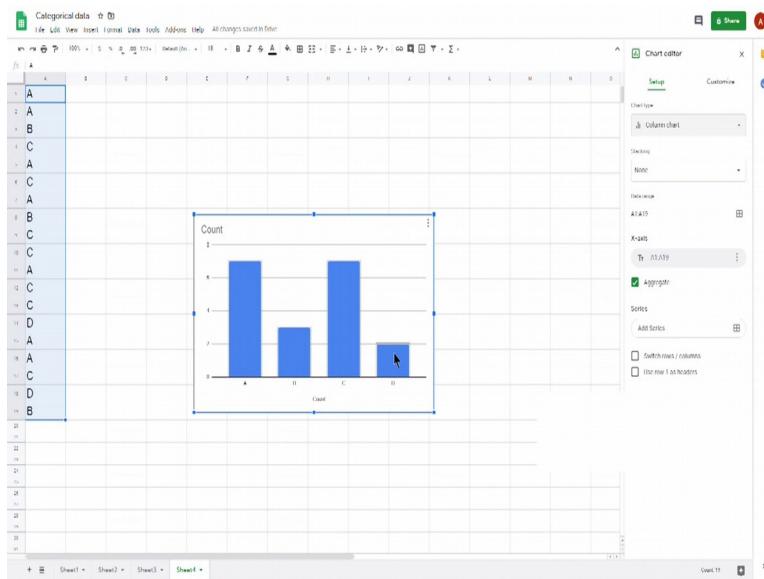
- ▶ If two or more categories tie for the highest frequency, the data are said to be bimodal (in the case of two) or multimodal (more than two).
- ▶ Let consider the example A,A,B,C,A,C,A,B,C,C, A,C,C,D,A,A,C,D,B

Both category "A" and "C" have highest frequency.

A woman in an orange sari is visible on the right side of the slide.

Now suppose, you have this data which is A, A, B, C, A so, I look at this other data where my data is A A B C A C A B C C.

(Refer Slide Time: 15:12)



So, this is my data set now and I just plot a column chart for this. Now what you look at this pie chart as you can see that both A and C are equally distributed or share of the pie is the same, which is 36.8%. Now this comes out much better in a bar chart where this is again 7 this is 7 B and C 2 and 1 respectively. But what is this info, what can we talk about this? So, when I have more than one category, in this case, I have A and C which have a frequency 7 each.

(Refer Slide Time: 15:55)

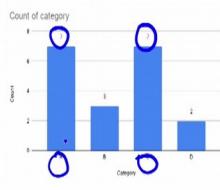
Statistics for Data Science -1
↳ Mode and Median



Bimodal and multimodal data

- If two or more categories tie for the highest frequency, the data are said to be bimodal (in the case of two) or multimodal (more than two).
- Let consider the example A,A,B,C,A,C,A,B,C,C,
A,C,C,D,A,A,C,D,B





► Both category "A" and "C" have highest frequency.




So, in a sense, both A and C have the highest frequency and I say the data is said to be bimodal. I repeat, if I have two categories or two or more categories that tie for the highest frequency as in this case, both A and C had a frequency of 7 each which is the highest frequency. Hence both A and C are having a tie for the highest frequency. My data is referred to as a by bimodal data.

If I have more than two categories, which are where there is a tie for the highest frequency, I refer to that data as multimodal data. So, this is about both Category A and C have the highest frequency.

(Refer Slide Time: 17:05)

Statistics for Data Science -1
└ Mode and Median

Median



▶ Ordinal data offer another summary, the median that is not available unless the data can be put into order.



Statistics for Data Science - I

I - Mode and Median

Median

- ▶ Ordinal data offer another summary, the median, that is not available unless the data can be put into order.

Definition

The **median** of an ordinal variable is the category of the middle observation of the sorted values.

- ▶ If there are an even number of observations, choose the category on either side of the middle of the sorted list as the median.

Now, another useful descriptive measure is what we refer to as a median. But a word of caution here is my data has to be ordered or ordinal in nature, I cannot define a median when I have a data, categorical data which is nominal. Again remember, small, medium, large, XL these are sizes of T shirts, this is ordinal data. A, B, C D if they referred to grades, this is again ordinal data because A is better than a B grade which is better than a C grade which is better than a D grade usually.

Now when you have ordinal data, I can offer another summary which is called the median. When we talk about mode for categorical data, we are counting that variable which appears the most number of times or that variable which has the highest frequency and that is what we refer to as the mode. Now, what is a median? Now, in order to compute a median, we first require the data to be ordered or ordinal data. So, unless the data can be put in some sort of an order, we cannot talk about the measure median, what is the measure median?

So, I can define the median of an ordinal variable to be that category of the middle observation of the sorted values. So, if there are even number of observations than the median could be either of the values of the middle, middle there be two middle observations, it could be either of those values, if it is odd then it is exactly the middle observation.

(Refer Slide Time: 19:16)

Statistics for Data Science - I
 I. Mode and Median

Example

$A B C D \quad n=15 < \text{Odd}$

<ul style="list-style-type: none"> ▶ Consider the grades of 15 students which is listed as <p>A,B,B,C,A,D,B,B,A,C, B,B,C,D,A</p> <p>1 1 2 1 2 1 3 1 3 2 6 6 3 2 4</p>	$A = 1$ $A = 2$ $A = 3$ $A = 4$ $B = 5$ $B = 6$ $B = 7$ $B = 8$ $B = 9$ $B = 10$ $C = 11$ $C = 12$ $C = 13$ $D = 14$ $D = 15$
--	---



So, how do we compute the median? So, now you consider the grades of 15 students, which is listed as given. So, I have a A. Now, if I am going to actually order this consider A is the highest grade B is the next highest C is the next highest and D is the next highest. I am going to have an order of this kind. I have an A I have how many A's I have 1, 2, 3, 4 I have four A's So, I can order it A, A, A, A. So, my first observation is in A second is again an A, and a third is an A fourth is an A that is a rank order given to these observations, then I have a B 1, 2, 3, 4, 5, and 6.

So, I write 6 B's ,1,2,3,4,5 and 6. So this is my fifth, 6,7,8,9, and 10. I have a C 1, 2, 3 a C, C, C, 11, 12, 13, I have a D. 2 D's. So 14, 15. So, what I have done here is we have listed the variables, what are the variables I have here I have A, I have B, I have C and I have D these are the variables. I have here. I have listed them in an order the order is A is better than B is better than C is better than D.

You could have chosen to have it in the other order also. Now once I have listed the variables in this order, you can see that the number of observations n equal to 15 is odd number that is what you can notice. So, if I look at what is that observation which divides this data set into exactly two halves, you can observe that this 8 observation here, has 7 above it 1, 2, 3, 4, 5, 6, 7. 1, 2, 3, 4, 5, 6, 7. So, this 8 observation which is corresponding to the variable B is the middle observation of the data set.

Hence, I can find out the ones I have ordered data, the median grade is that value which is associated with the 8 observation as given here and that is B.

(Refer Slide Time: 22:10)



Example

- ▶ Consider the grades of 15 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D,A
 - ▶ The ordered data is A,A,A,A,B,B,B,B,B,C,C,C,D,D
 - ▶ The median grade is the category associated with the 8 observation which is "B".
- ▶ Consider the grades of 14 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D
 - ▶ The ordered data is A,A,A,B,B,B,B,B,C,C,C,D,D
 - ▶ The median grade is the category associated with the 7 or 8 observation which is "B".

(□) (□) (□) (□) (□) (□) 44 / 47

Now suppose I consider the grades of 14 students, when I have a grade of 14 students may n equal to 14. I can repeat the same exercise, and I can write it as I have 1, 2, 3 A's, so, A, A, A have 1, 2, 3, 4, 5, 6 B's, B, B, B. I have 1, 2, 3 C's and I have two D's. I again write an order. In this case, you can see the following is that I have n which is equal to 14. So, there is not n observation which can exactly have equal number of observations above it and below it.

So in this case, if I look at the seventh and eighth observations together, you can see it has 6 above it and 6 below it. So again, my median is again B, which is either the seventh observation, or the eighth observation, which in this case is B. Now, in case instead of the eight observation instead of a B I had a C, the median would have been either B, or C. So, this is how you compute a median of a categorical data that is ordinal. Why would you compute a median for categorical data?

The mode in a sense gives you the length of the longest bar in a bar chart or the largest pie in the pie chart because it gives you the that category or that variable which has the highest frequency. The case of a median when you have ordered data in a sense, for example, if this ABCD with sizes, then you would have said that this particular size particularly divides the data into two halves.

So, median in some sense gives you a measure of central tendency. We will look at this measure in greater detail when we look into measures for numerical data.

(Refer Slide Time: 24:45)

Statistics for Data Science - I

Mode and Median



Example

$N=15$

A	4	A	1
B	6	A	2
C	3	A	3
D	2	B	4
		B	5
		B	6
		B	7
		B	8
		B	9
		B	10
		C	11
		C	12
		C	13
		D	14
		D	15

► Consider the grades of 15 students which is listed as
~~K, B, B, 1, S, A, D, B, B, A, 7, 7, F, F, D, F~~

(口)(@)(三)(三) 三 丙戌年 15/17

Statistics for Data Science - I
Mode and Median

Example



- ▶ Consider the grades of 15 students which is listed as A,B,B,C,A,D,B,B,A,C, B,B,C,D,A
 - ▶ The ordered data is A,A,A,A,B,B,B,B,B,C,C,C,D,D
 - ▶ The median grade is the category associated with the 8 observation which is "B".
 - ▶ The most common grade is "B", hence mode is "B"
 - ▶ In this example both mode and median are the same.

« □ » « ⌂ » « ⌃ » « ⌄ » ⌁ ⌂ ⌃ ⌄ 45 / 47

So, again, if you consider the grades of these 15 students again, I go back to the example here how again listed the grades of the 15 students. So, again, what do we have I have n equal to 15 I have listed the grades. So let us again count how many A's do I have I have 1, I have a 2, I have a 3, I have a 4.

So, I have a count of 4 for A, I have a count of 1, 2, 3, 4, 5, 6 a count of B, I have 6, I have C's 1, 2, 3 and I have a D, which is 2. So, this is again the same example I have a A, A ,A, A I have a B, B, B, B, B, B. I have a C, C, C and a D, D rank them 1, 2, 3 ,4, 5, 6, 7, 8, 9 10, 11, 12, 13, 14

and 15 and I see the median is B. The mode is also B because B is that category which appears the largest frequency. So in this example, the median is B, the mode is also B. So the natural question to ask is, will the median and mode in any data set be the same. The answer is no.

(Refer Slide Time: 26:18)

Statistics for Data Science -1

1. Mode and Median



Example

$n=15$

	A	B	C	D
A	1	2	3	4
B	1	2	3	4
C	1	2	3	4
D	1	2	3	4

► Consider the grades of 15 students which is listed as
~~A, B, B, C, D, A, B, C, D, A, B, C, D, E, F, G, H, I, K, L, X~~

	B	C	D	E	F	G	H	I	K	L	X
B	1	2	3	4	5	6	7	8	9	10	11
C	1	2	3	4	5	6	7	8	9	10	11
D	1	2	3	4	5	6	7	8	9	10	11
E	1	2	3	4	5	6	7	8	9	10	11
F	1	2	3	4	5	6	7	8	9	10	11
G	1	2	3	4	5	6	7	8	9	10	11
H	1	2	3	4	5	6	7	8	9	10	11
I	1	2	3	4	5	6	7	8	9	10	11
K	1	2	3	4	5	6	7	8	9	10	11
L	1	2	3	4	5	6	7	8	9	10	11
X	1	2	3	4	5	6	7	8	9	10	11

Example

- ▶ Consider the grades of 15 students which is listed as A,B,B,C,A,D,A,B,A,C,B,A,C,D,A
 - ▶ The ordered data is A,A,A,A,A,B,B,B,C,C,C,D,D
 - ▶ The median grade is the category associated with the 8 observation which is "B".
 - ▶ The most common grade is "A", hence mode is "A"
 - ▶ In this example both mode and median are the different.

For example, consider the given data set. Again I repeat the exercise, I have again 4 categories A, B, C, D, the total number of observations is again 15. Again, let us count A, A 1, 2, 3, 4, 5, 6 A is 6, B is 1, 2, 3, 4, C is 1, 2, 3. And D is 2. Now let us look at the median I have A, A, A, A, A, A, B, B, B, C, C, C, D, D. So, I have ranking 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

again the median for this data set is B, whereas the mode is A because it appears with the highest frequency and the highest frequency is given by 6.

So this is an example of a data set where the median grade is B, whereas the mode grade is A. Sometimes these summaries are helpful in summarizing the performance of a class by relating or by giving a numerical descriptive measure instead of graphical measure in terms of mode and median.

(Refer Slide Time: 27:47)



Sectional summary

- ▶ The mode of a categorical variable is the most common category.
- ▶ The median of an ordinal variable is the category of the middle observation of the sorted values.



So, what we have learned so far is when we come to numerical descriptive measures, in case of categorical data, there are two important measures. The first measure as what we refer to as a mode, which just gives you which is the most common category. The second measure is what we refer to as the median, which is the middle observation of sorted values, the mode works for nominal data, but to calculate the median, you need the data to be ordinal. So, with this we come to an end to with the module on describing categorical variables.

(Refer Slide Time: 28:31)



Summary

1. Tabulate data: frequency and relative frequency.
2. Charts of categorical data
 - 2.1 Pie charts
 - 2.2 Bar charts and Pareto charts
3. Best practices and misleading graphs
 - 3.1 Label your data.
 - 3.2 Dealing with multiple categories.
 - 3.3 Area principle
 - 3.4 Misleading graphs
 - 3.4.1 Decorated graphs
 - 3.4.2 Truncated graphs.
 - 3.4.3 Round-off errors.
4. Descriptive measures
 - 4.1 Mode.
 - 4.2 Median for ordinal data.



So, what should we have learned here is we started with tabulating data. What do we understand by tabulating data? We first identify what are the number of categories of my particular variable. If there how

many categories I state all these categories A, B, C, D, How many of our categories I have a list down all the categories. When I look at a tabulating or a frequency table, I have categories, I have the frequency.

Frequency is just the count of each category, I define what is the relative frequency, the relative frequency is the frequency by total count or total number of observations. So, this is frequency by total number of observations. If I had to summarize it using a pie chart where the question or purpose is to look at what is it as a composition of the whole, I plot the relative frequencies in terms of a pie chart, I can also plot it through a frequency which is called a bar chart.

A Pareto chart is where my categories are actually arranged in either decreasing or increasing order. When we have either a pie chart or a bar chart, labeling the data or annotating it, having appropriate titles labeling your access all of them are extremely important. Do not ignore categories when you deal with multiple categories that is many categories. Instead club all the categories which have very few counts into one because they convey information we saw an

example where the multiple categories which had very low counts actually conveyed a lot of information.

Decorated graphs avoid if you can, do not use truncated grass which truncate the baseline and use An artificial baseline it could be very misleading, be aware of round off errors because round off errors could also not add up to your totals. Finally, we looked at descriptive measures. We looked at mode, mode is only for nominal when you have nominal data. The only possible descriptive measure is that for mode, when we have ordinal data we can have both mode than median.

We looked at an example where mod was equal to the median and we also looked at an example where the mode was actually lesser than the median. So, what where do we stand now with respect to the course with respect to the course we have now we understand what is our type of data we know how to categorize these data.

So given a variable, first given a data I can identify what are my variables I can identify what are my cases. I know how to tabulate my data, now given my data I can actually classify them into categorical or numerical. And given a categorical variable I know how to tabulate my categorical variable and summarize my categorical variable using a pie chart or by bar chart and to also give a descriptive measure in terms of a mode and median.

Statistics for Data Science - 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture – 2.1
Problems Charts and Tables

(Refer Slide Time: 00:14)

Statistics for Data Science - 1
Week 2 Tutorial Questions
Describing categorical data - one variable



Syllabus covered:

- Organizing and graphing categorical data.
- Create frequency tables for tabulated data.
- Choosing an appropriate graphical technique for displaying data.
- Discuss about misleading graphs.

Hello statistics students. In this tutorial we are going to do problems based on the concepts covered in week 2. So, this is the syllabus and let us begin with our first question.

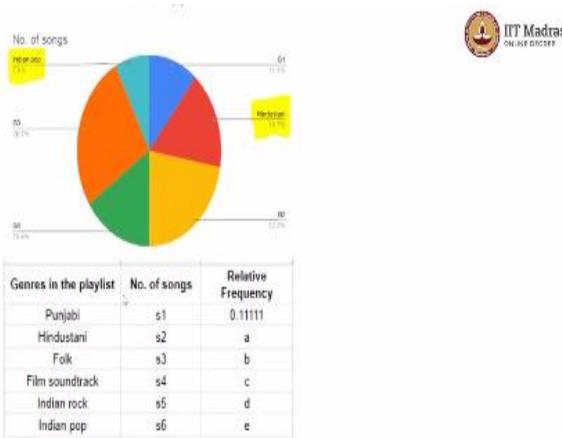
(Refer Slide Time: 00:31)



Here they are saying given the information here in all these figures. So, there are three figures there is a bar chart, a pie chart and the frequency table. Using this information answer the following questions. So, what is this information? This is essentially to do with songs and

genres in a playlist. So, there are these genres here which is Punjabi, Hindustani, folk film soundtrack, Indian rock and Indian pop. And the number of songs for each of these are given actually not for all of them. Hindustani and Indian pop are empty, there is nothing over there. So, this is an incomplete bar graph.

(Refer Slide Time: 01:26)



- Find the number of songs corresponding to Hindustani genre?
- Find the number of songs corresponding to Indian Pop genre?
- Which is the modal genre in the playlist?

Now, let us look at the pie chart. The pie chart is presumably for the same data, we have Indian pop and Hindustani given here with percentages. The rest of them are unnamed. We know the percentages, but they just named G_1, G_2, G_3, G_4 and lastly we have this frequency table which looks mostly unfilled, just given the number of songs as variables and relative frequency also variables except this one particular relative frequency for the Punjabi genre.

So, this number 0.11111 if you observe we can actually reduce it to a nice $\frac{P}{Q}$ form in rational if you are following in the Math course. Let us say this is x then that x is essentially 0.111111, so on. If I multiply this by 10 I would get 1.111111 so on. Now, let us subtract this whole thing from the previous one. What do we get? We have $9x = 1$ that would imply $x = \frac{1}{9}$.

So, this value here the one relative frequency as given it is $\frac{1}{9}$, but we do not know if this is going to be useful for us in this question, but you can reduce it to $\frac{1}{9}$, if you like. Now, let us go on to actual questions. Find the number of songs corresponding to the Hindustani genre. Let us see here, as expected the Hindustani one it is empty, we do not know what is Hindustani genre songs from this particular graph, let us look at the next one.

So, we know it is 16.7 % of the total the problem is we do not know what the total is, so how do we crack this. Let us look at further more data. Indian pop is also empty and this value is not really going to help us now, but this value are $\frac{1}{9}$ is going to come and help us here the Punjabi songs we know there are 10 of them because this is where the bar graph is ending.

So, we know that one-ninth of the songs are 10 songs. So, total number of songs are n , then we know that $\frac{1}{9} \times n = 10$ and that implies $n = 90$ songs overall and we know now that Hindustani is 16.7% of 90 songs. So, let us calculate that this is essentially then $\frac{16.7}{100}$ it is very likely that the 16.7 is a approximate value, rounded off value. So, anyway let us find out what is happening here we get. 9×1.67 which is coming out to be 15.03.

So, we know that number of songs has to be an integer so this has to be essentially 15 songs. So, if we have to represent this in our bar graph, we would get something like this may be of course that is rough. Moving on, find the number of songs corresponding to the Indian pop genre. So, we have Indian pop to be 7.8%. So, again we have 7.8% of 90 songs. Let us calculate it here you would get $\frac{7.8}{100} \times 90$ which gives us 0.78×9 which is 7.02.

So, this must be rounded off to 7 songs. So, going back to our bar graph that probably comes still somewhere here so it looks something like this which is a modal genre in the playlist. We know that modal genre is the genre with most frequency. So, here that would be Indian rock evidently Indian rock has the maximum. So, this must be this value must be about 24 it is a little less than 25, but it is definitely more than half of this particular line considerably more than half of this. So, it cannot be 23 either, so it has to be 24.

So, Indian rock is our modal genre so this is Indian rock. Finally we are being asked find the appropriate genre from the playlist that corresponds to $G1, G2, G3, G4$ which are what are here. We know that $G1$ has 11.1% so 11.1% of 90 would be again $\frac{11.1}{100} \times 90$ which is basically 9.99 so this must be roughly 10 songs. So, let us see which genre has 10 songs it is Punjabi.

So, $G1$ must be Punjabi let us look at $G2$. $G2$ is twice the percentage of $G1$ so this must be 20 songs and which has 20 songs here, folk. So, $G2$ must be folk and then look at $G3, 26.7\%$. What is 26.7% of 90? So, $\frac{26.7}{100} \times 90$, 0, 0 cancels off we are getting 2.67×9 . What would that

be? This is essentially 24.03. So, roughly 24 songs and we have seen earlier that 24 songs is the number for our modal genre which is Indian rock.

So, G_3 is certainly Indian rock and lastly G_4 is whatever is left and what is left here, everything else is done, Indian pop is done, Punjabi is done, Hindustani is done, Indian rock is done, folk is also done, so G_4 definitely to be film soundtrack. So, G_1 was Punjabi, G_2 was folk, G_3 is Indian rock and lastly G_4 is film soundtrack.

(Refer Slide Time: 10:28)

Genre	Number of Songs	Relative Frequency
Punjabi	10	0.111111
Hindustani	15	0.166667
Folk	20	0.222222
Film Soundtrack	14	0.155556
Indian Rock	24	0.266667
Indian Pop	7	0.077777

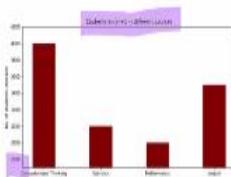
And so this is the final frequency table along with the corresponding relative frequencies. If you add up all these numbers you will get $10 + 15$ is 25, $25 + 20$ is 45, $45 + 14$ is 59, $59 + 24$ is 83 and $+ 7$ gives us 90, which is what we expect and likewise if you sum the relative frequencies you will 1. Thank you.

Statistics for Data Science - 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture – 2.2
Problems Misleading Graphs

(Refer Slide Time: 00:14)



The figure given below shows the number of students enrolled in different courses. Why is this graph misleading?

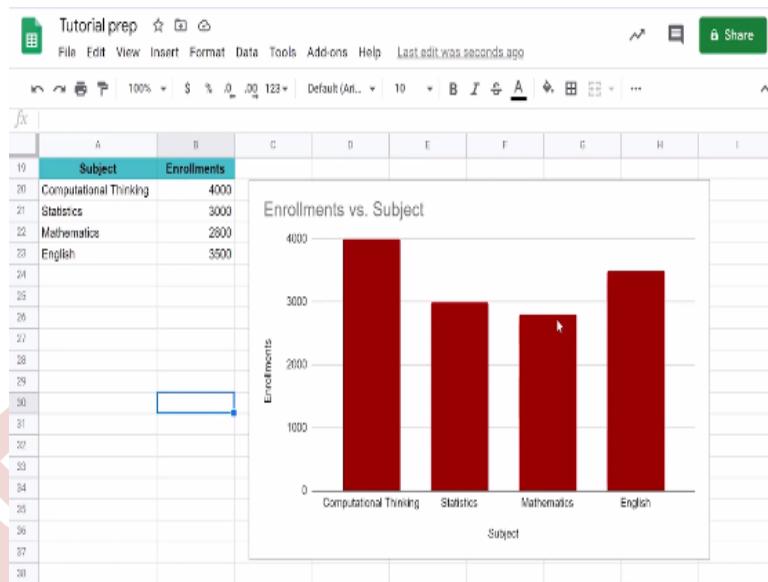


- Graphed incorrectly
- Not misleading
- Does not start at zero
- The width of the bin is uniform

In this video let us look at two misleading graph questions. The figures given below shows the number of students enrolled in different courses. Why is this graph misleading? So, students enrolled in different courses. This is the number of students and this is the courses offered. So, we have our four courses over here. So, is this graphed incorrectly that is the vague question and we do not even know about the original data. So, we cannot really consider to be a proper option one way or the other.

Is the graph not misleading we have to find out. Does not start at zero, yes, that fixes it. Clearly we are starting somewhere around 2400 because each of these units seems to be about 200 students and the one below 2600 must be 2400. So, we are not starting from zero this is the issue. The width of the bin is uniform, but there is nothing wrong with the width of the bin being uniform. So, that must be your answer.

(Refer Slide Time: 01:27)



Now, this is the actual data 4000, 3000, 2800 and 3500 and from this data let us select it. We will try to make a bar graph which is more representative of this data. So, we selected this data and now we go to insert chart and showing as a column chart if you want to customize we could always go and change the colors so this is the color that was used in the question. So, this is what our graph is expected to look like which is actually all of them look roughly the same when you compare with the baseline as zero. But over in the question the baseline was somewhere here which made the 2800 much smaller than what it actually was.

(Refer Slide Time: 02:23)

A survey was conducted to determine what food items would be served at the farewell party. Why is this graph misleading?

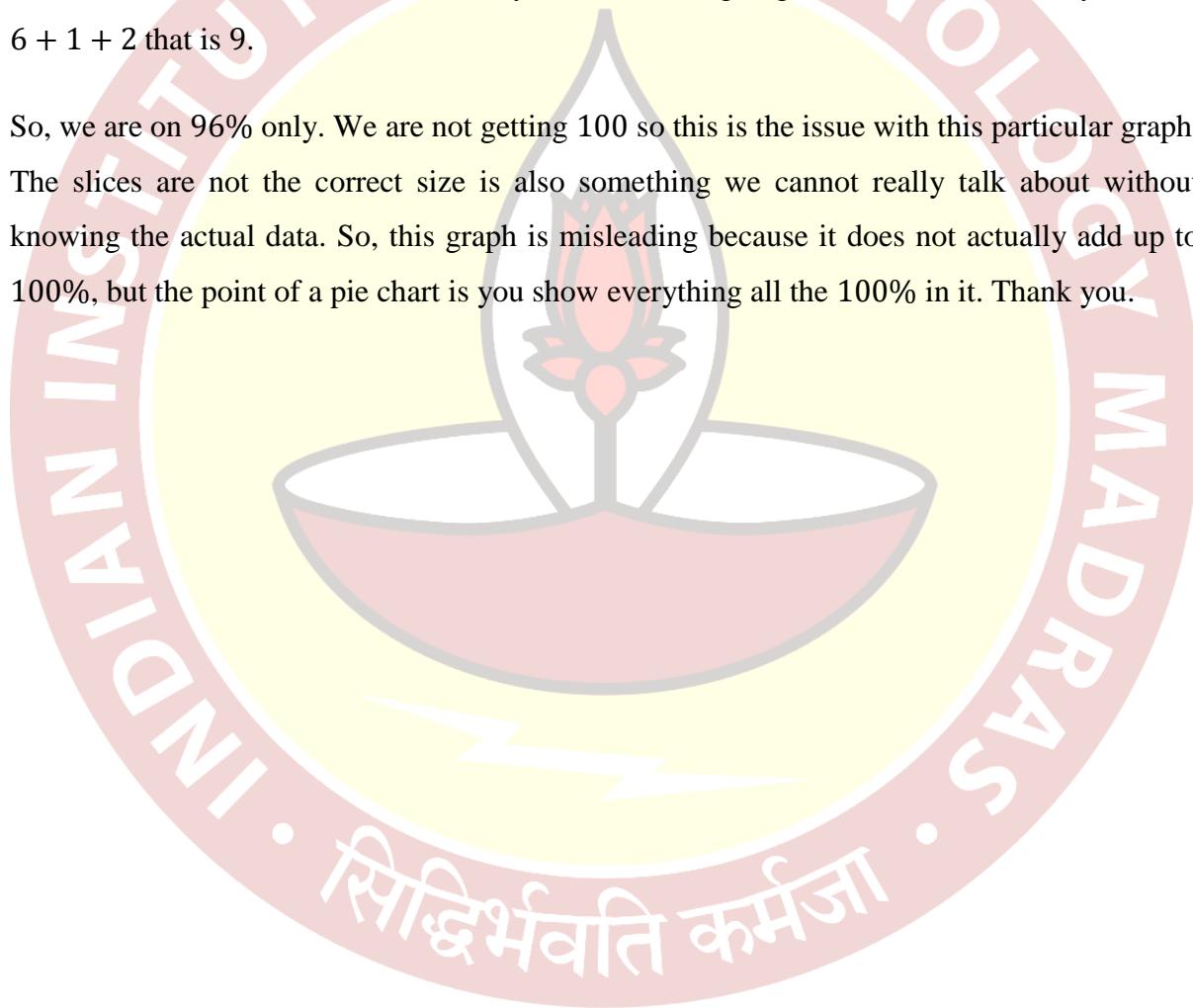


- Not misleading.
- The slices are the wrong colour.
- The percentages do not add up to 100.
- The slices are not the correct size.

In this pie chart now let us look at this. A survey was conducted to determine what food items would be served at the farewell party why is this graph misleading. So, we have these four food items chilly potato, spring rolls, cheese corn balls and Nachos. We have these percentages associated with them. This graph is not misleading we do not yet we have to see. The slices are the wrong colour.

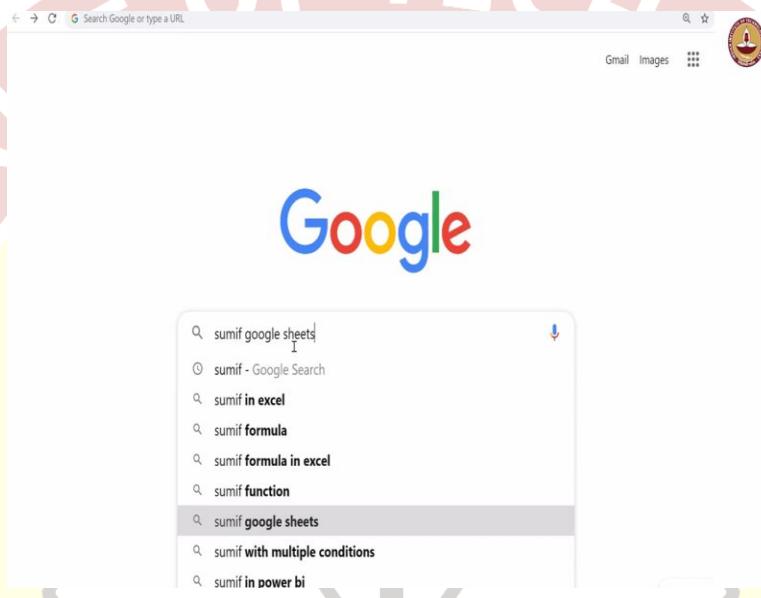
There was no way to say that unless we have the real data we cannot really say that for now this pie chart could very well be fine. The percentages do not add up to 100. Now, let us verify this, this is important so I have $29\% + 28\% + 15\% + 24\%$ and what we have we got, 9 + 8 is 17, 17 + 5 is 22, 22 + 4 is 26. So, yeah this is not going to be 100, but to carry over so 6 + 1 + 2 that is 9.

So, we are on 96% only. We are not getting 100 so this is the issue with this particular graph. The slices are not the correct size is also something we cannot really talk about without knowing the actual data. So, this graph is misleading because it does not actually add up to 100%, but the point of a pie chart is you show everything all the 100% in it. Thank you.



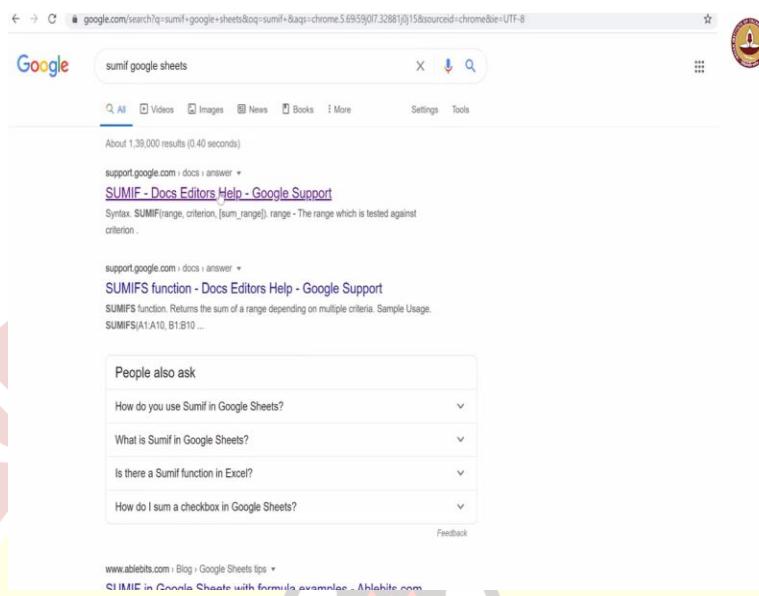
Statistics for Data Science
Professor. Usha Mohan
Prathyush P (Support Team)
Department of Management Studies
Indian Institute of Technology, Madras
Tutorial - 3
SUMIF in Google Sheets

(Refer Slide Time: 00:14)



Hello Statistics students. In this tutorial, we will look at another very useful function in Google Sheets, which is SUMIF. So, we search for SUMIF Google Sheets.

(Refer Slide Time: 00:28)



Google sumif google sheets

About 1,39,000 results (0.40 seconds)

[support.google.com / docs / answer / 3093583?hl=en](#)

SUMIF - Docs Editors Help - Google Support

Syntax: `SUMIF(range, criterion, [sum_range])`, range - The range which is tested against criterion.

[support.google.com / docs / answer / 3093583?hl=en](#)

SUMIFS function - Docs Editors Help - Google Support

SUMIFS function. Returns the sum of a range depending on multiple criteria. Sample Usage. `SUMIFS(A1:A10, B1:B10 ...)`

People also ask

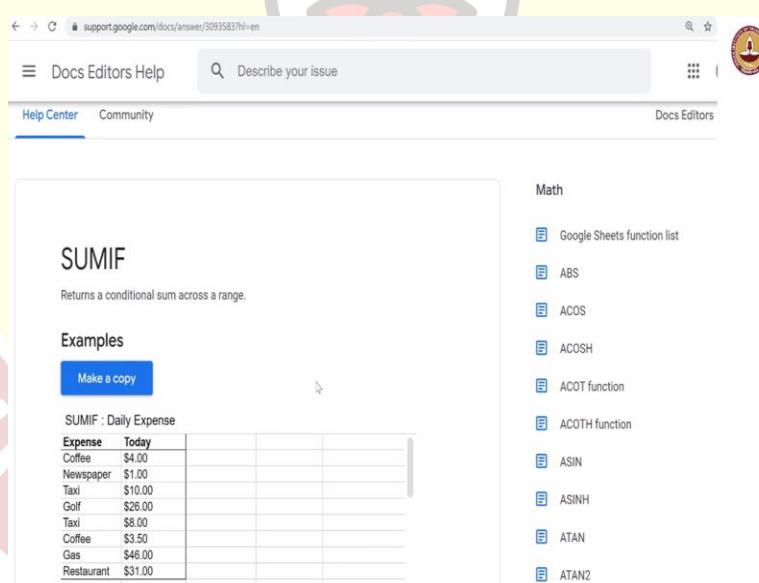
- How do you use Sumif in Google Sheets?
- What is Sumif in Google Sheets?
- Is there a Sumif function in Excel?
- How do I sum a checkbox in Google Sheets?

Feedback

www.ablebits.com / Blog - Google Sheets tips •
CLIMIC in Google Sheets with formula examples - Ablebits.com

And here we again go to the Doc Editors help.

(Refer Slide Time: 00:34)



support.google.com/docs/answer/3093583?hl=en

Docs Editors Help

Describe your issue

Help Center Community Docs Editors

SUMIF

Returns a conditional sum across a range.

Examples

Make a copy

SUMIF : Daily Expense

Expense	Today
Coffee	\$4.00
Newspaper	\$1.00
Taxi	\$10.00
Golf	\$26.00
Taxi	\$8.00
Coffee	\$3.50
Gas	\$48.00
Restaurant	\$31.00

Math

- Google Sheets function list
- ABS
- ACOS
- ACOSH
- ACOT function
- ACOTH function
- ASIN
- ASINH
- ATAN
- ATAN2

And here we are in this page, which explains SUMIF to us.

(Refer Slide Time: 00:41)

The screenshot shows a Google Support article page for the SUMIF function. The 'Sample Usage' section contains two examples: `SUMIF(A1:A10, ">20")` and `SUMIF(A1:A10, "Paid", B1:B10)`. The 'Syntax' section defines the formula as `SUMIF(range, criterion, [sum_range])` and provides detailed explanations for each parameter. The 'range' parameter is described as the range tested against the criterion, which can contain wildcards like '?' or '*' or the tilde (~) character. The 'criterion' parameter is the pattern or test applied to the range. The 'sum_range' parameter is the range to sum if the criterion is met. A sidebar on the right lists related functions such as CEILING.MATH, CEILING.PRECISE, COMBIN, COMBINA, COS, COSH, COT, COTH, COUNTBLANK, COUNTIF, COUNTIFS, COUNTUNIQUE, and CSC.

And the syntax is SUMIF and the first parameter is a range, which is the range over which our criterion is going to be tested. And the second parameter is the criterion that we want to test the pattern or test to be applied to range and then the sum range is the range that needs to be summed if that criterion is matched. So, what we are doing in sumif is this is basically summing with an IF. So, if this criterion is matched on this range, then we do the sum on this sum range. So, the example for this is sum range.

(Refer Slide Time: 01:30)

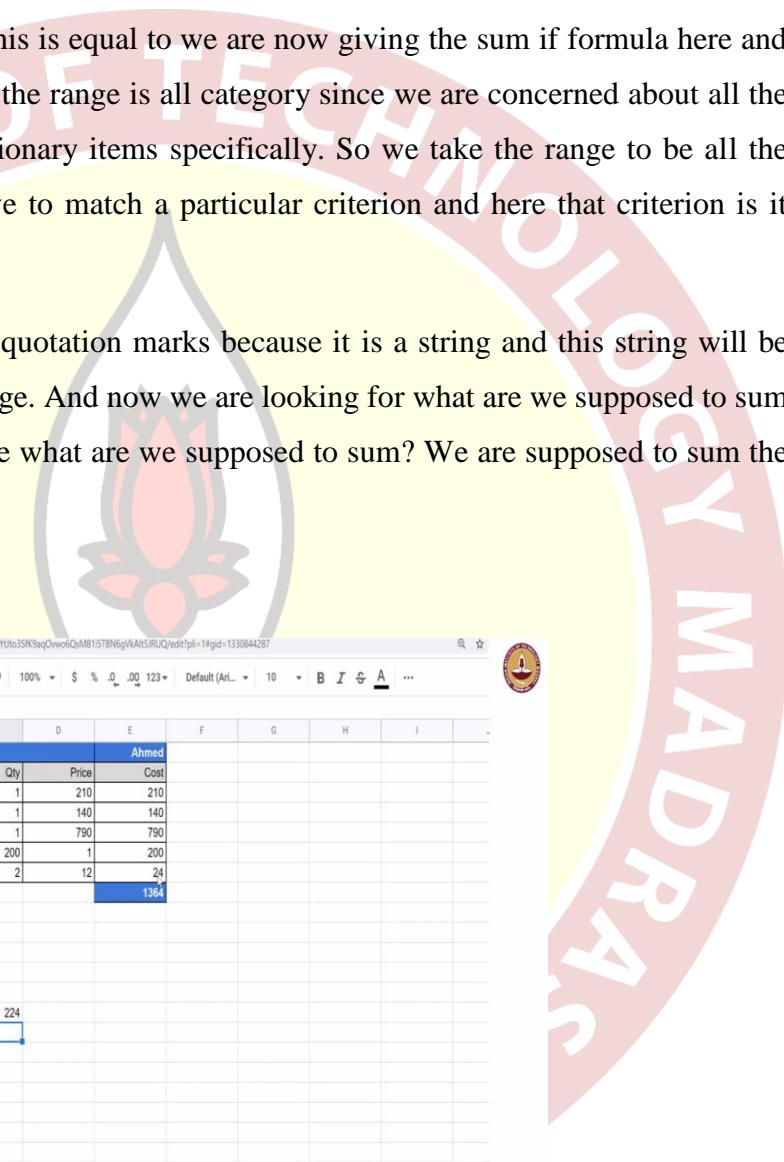
The screenshot shows a Google Sheets spreadsheet with data in columns A through E. Row 1 contains headers: 'Item', 'Category', 'Qty', 'Price', and 'Cost'. Rows 2 through 7 contain data: Earphones (Electronics, Qty 1, Price 210, Cost 210), Phone cover (Accessories, Qty 1, Price 140, Cost 140), Dongle (Electronics, Qty 1, Price 790, Cost 790), A4 sheets (Stationery, Qty 200, Price 1, Cost 200), and Ball Pens (Stationery, Qty 2, Price 12, Cost 24). The total cost for the 'Stationery' category is calculated in cell E13 using the formula `=sumif(B3:B7, "Stationery", E3:E7)`. The formula is also shown in the formula bar above cell E13.

You probably recognize this as a Shopping bill data set card from the computational thinking course. Here, let us say we want to add all the stationery items Cost that would be $200 + 24$, we should get 224. So once again, this is an example. In reality, you are likely to have really large data sets, where you can not personally physically go through all of them and do these sums.

So, how we do it in these cases is this is equal to we are now giving the sum if formula here and the first parameter is the range. So, the range is all category since we are concerned about all the Categories, we are looking for stationary items specifically. So we take the range to be all the categories and these categories have to match a particular criterion and here that criterion is it should be stationary.

So, I am putting it in these double quotation marks because it is a string and this string will be matched against the cells in that range. And now we are looking for what are we supposed to sum once we find stationary in this range what are we supposed to sum? We are supposed to sum the cost so this is the summing range.

(Refer Slide Time: 03:05)

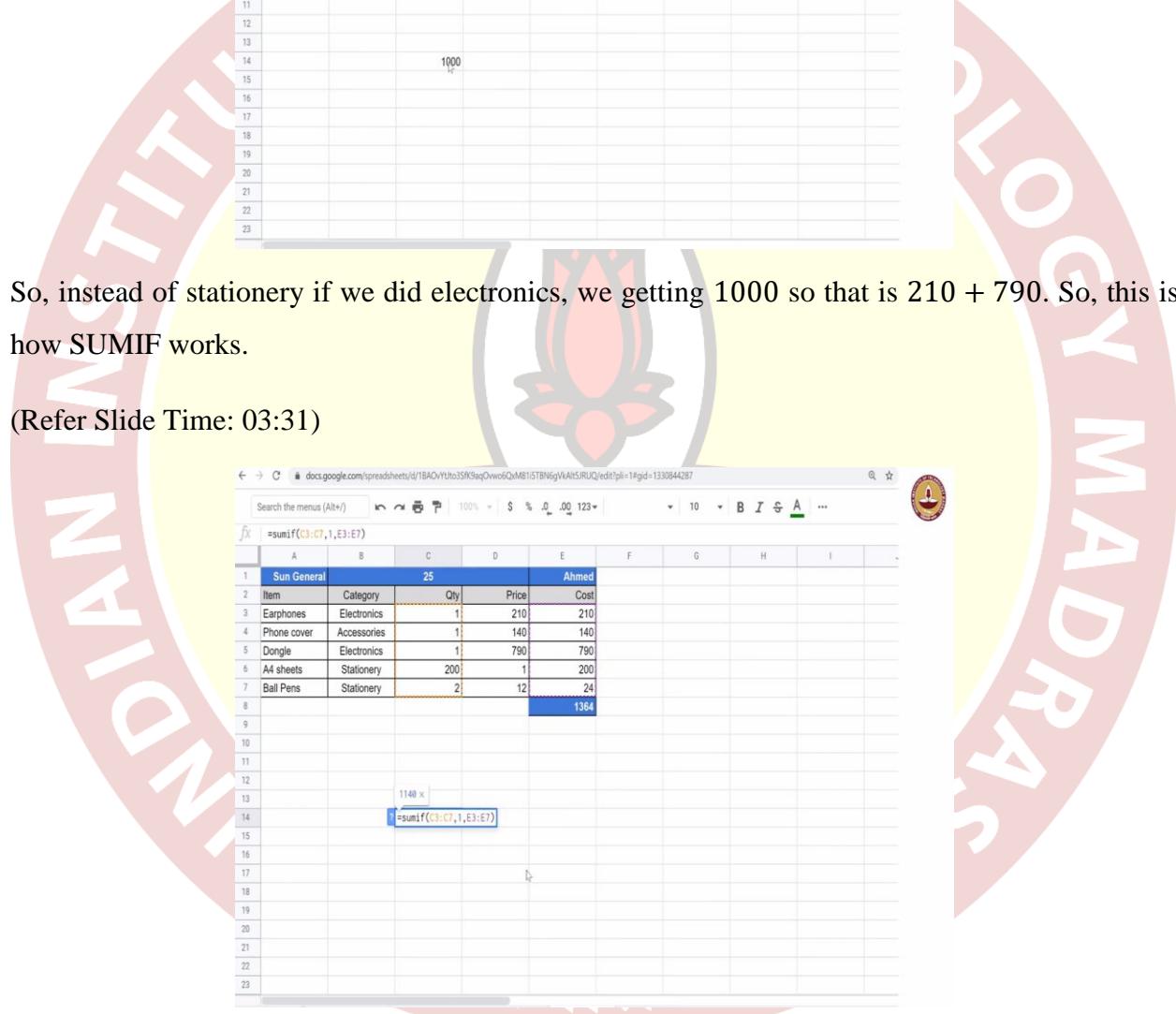


A screenshot of a Google Sheets spreadsheet titled "Sun General". The data is as follows:

	A	B	C	D	E	F	G	H	I
1	Sun General	25			Ahmed				
2	Item	Category	Qty	Price	Cost				
3	Earphones	Electronics	1	210	210				
4	Phone cover	Accessories	1	140	140				
5	Dongle	Electronics	1	790	790				
6	A4 sheets	Stationery	200	1	200				
7	Ball Pens	Stationery	2	12	24				
8					1364				
9									
10									
11									
12									
13									
14				224					
15									
16									
17									
18									
19									
20									
21									
22									
23									

And then we press enter here we are with 224 so that is $200 + 24$.

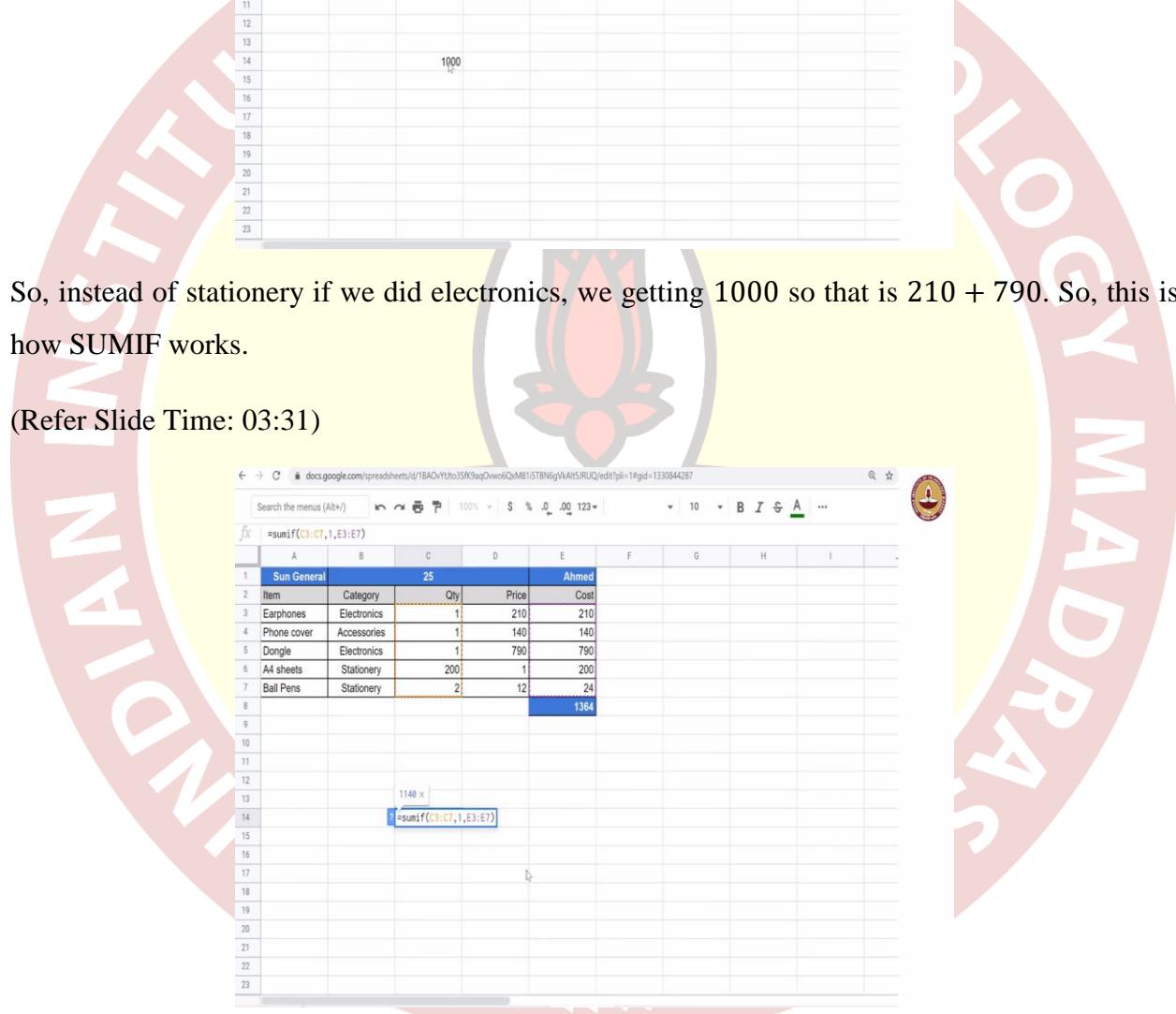
(Refer Slide Time: 03:20)



A	B	C	D	E	F	G	H	I
1	Sun General	25		Ahmed				
2	Item	Category	Qty	Price	Cost			
3	Earphones	Electronics	1	210	210			
4	Phone cover	Accessories	1	140	140			
5	Dongle	Electronics	1	790	790			
6	A4 sheets	Stationery	200	1	200			
7	Ball Pens	Stationery	2	12	24			
8					1364			
9								
10								
11								
12								
13								
14					1000			
15								
16								
17								
18								
19								
20								
21								
22								
23								

So, instead of stationery if we did electronics, we getting 1000 so that is $210 + 790$. So, this is how SUMIF works.

(Refer Slide Time: 03:31)



A	B	C	D	E	F	G	H	I
1	Sun General	25		Ahmed				
2	Item	Category	Qty	Price	Cost			
3	Earphones	Electronics	1	210	210			
4	Phone cover	Accessories	1	140	140			
5	Dongle	Electronics	1	790	790			
6	A4 sheets	Stationery	200	1	200			
7	Ball Pens	Stationery	2	12	24			
8					1364			
9								
10								
11								
12								
13								
14					1140	x		
15								
16								
17								
18								
19								
20								
21								
22								
23								

Here we did the criterion matching against text instead, if we were to do it against numbers, for example, if we wanted to look at quantities that are purchased as single item, so the quantity should be 1. So here the, the range that we are matching the criterion against is this and the criterion we

are matching is whether it is 1 so now it is a number I do not need to use the quotation marks I can directly put 1 and Enter.

(Refer Slide Time: 04:02)

A screenshot of a Google Sheets document showing a table of purchases. The table has columns for Item, Category, Qty, Price, and Cost. Row 1 is a header with 'Sun General' in A1, '25' in B1, and 'Ahmed' in E1. Rows 2 through 7 list items: Earphones (Qty 1, Price 210, Cost 210), Phone cover (Qty 1, Price 140, Cost 140), Dongle (Qty 1, Price 790, Cost 790), A4 sheets (Qty 200, Price 1, Cost 200), and Ball Pens (Qty 2, Price 12, Cost 24). The total cost is 1364, which is highlighted in blue. In cell C15, there is a formula bar showing '=sumif(C3:C7,>1,E3:E7)'.

Sun General		25	Ahmed	
Item	Category	Qty	Price	Cost
Earphones	Electronics	1	210	210
Phone cover	Accessories	1	140	140
Dongle	Electronics	1	790	790
A4 sheets	Stationery	200	1	200
Ball Pens	Stationery	2	12	24
				1364
			1140	

And so we have $210 + 140 + 790$, $210 + 790$ it is already 1000 we had seen additionally 140 gives us 1140.

(Refer Slide Time: 04:15)

A screenshot of a Google Sheets document showing the same table of purchases. The formula in cell C15 has been changed to '=sumif(C3:C7,>1,E3:E7)', as shown in the formula bar. The rest of the table and its contents remain the same as in the previous slide.

Sun General		25	Ahmed	
Item	Category	Qty	Price	Cost
Earphones	Electronics	1	210	210
Phone cover	Accessories	1	140	140
Dongle	Electronics	1	790	790
A4 sheets	Stationery	200	1	200
Ball Pens	Stationery	2	12	24
				1364
			=sumif(C3:C7,>1,E3:E7)	

We could do other kinds of evaluations on numbers. For instance, we can look for items that were purchased more than 1 at a time. So, the quantity should be greater than 1.

(Refer Slide Time: 04:31)

A screenshot of a Google Sheets spreadsheet. The formula bar shows '=sumif(C3:C7,>1,E3:E7)'. The spreadsheet contains the following data:

	A	B	C	D	E	F	G	H	I
1	Sun General	25			Ahmed				
2	Item	Category	Qty	Price	Cost				
3	Earphones	Electronics	1	210	210				
4	Phone cover	Accessories	1	140	140				
5	Dongle	Electronics	1	790	790				
6	A4 sheets	Stationery	200	1	200				
7	Ball Pens	Stationery	2	12	24				
8					1364				
9									
10									
11									
12									
13									
14				#ERROR!	Error				
15					Formula parse error.				
16									
17									
18									
19									
20									
21									
22									
23									

And here we have an error because greater than 1 is not a number.

(Refer Slide Time: 04:34)

A screenshot of a Google Sheets spreadsheet. The formula bar shows '=sumif(C3:C7,">1",E3:E7)'. The spreadsheet contains the same data as the previous slide:

	A	B	C	D	E	F	G	H	I
1	Sun General	25			Ahmed				
2	Item	Category	Qty	Price	Cost				
3	Earphones	Electronics	1	210	210				
4	Phone cover	Accessories	1	140	140				
5	Dongle	Electronics	1	790	790				
6	A4 sheets	Stationery	200	1	200				
7	Ball Pens	Stationery	2	12	24				
8					1364				
9									
10									
11									
12									
13				224 x					
14				=sumif(C3:C7,">1",E3:E7)					
15									
16									
17									
18									
19									
20									
21									
22									
23									

So, formula so criterion we are testing again. So, as a formula, this has to be put in quotation marks.

And now we press enter and here we getting 224 which is these two 200 + 24.

(Refer Slide Time: 04:51)

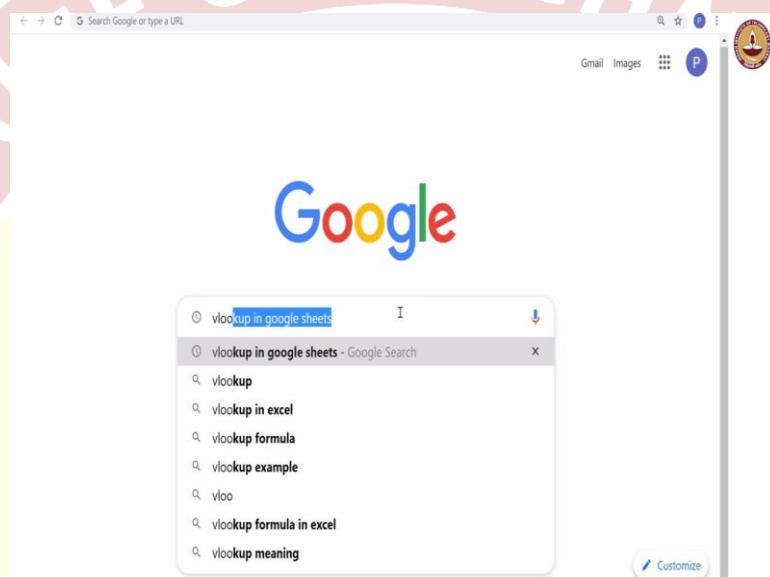
A screenshot of a Google Sheets document showing a table of inventory items. The table has columns for Item, Category, Qty, Price, and Cost. Row 25 is highlighted in blue, and the cell containing the value '1364' is also highlighted in blue. The formula bar at the top shows the formula =SUMIF(A2:A7,C25,D2:D7). The URL in the address bar is https://docs.google.com/spreadsheets/d/1BAQvYUto35R3aqQwofQ/M815TBn6gViAhSIUQ/edit?usp=sharing.

Item	Category	Qty	Price	Cost
Earphones	Electronics	1	210	210
Phone cover	Accessories	1	140	140
Dongle	Electronics	1	790	790
A4 sheets	Stationery	200	1	200
Ball Pens	Stationery	2	12	24
				1364
				224

Because these are the only quantities which are greater than 1. So in this way, we can use SUMIF to specifically sum certain cells based on some criterion which is matched against a range. Thank you.

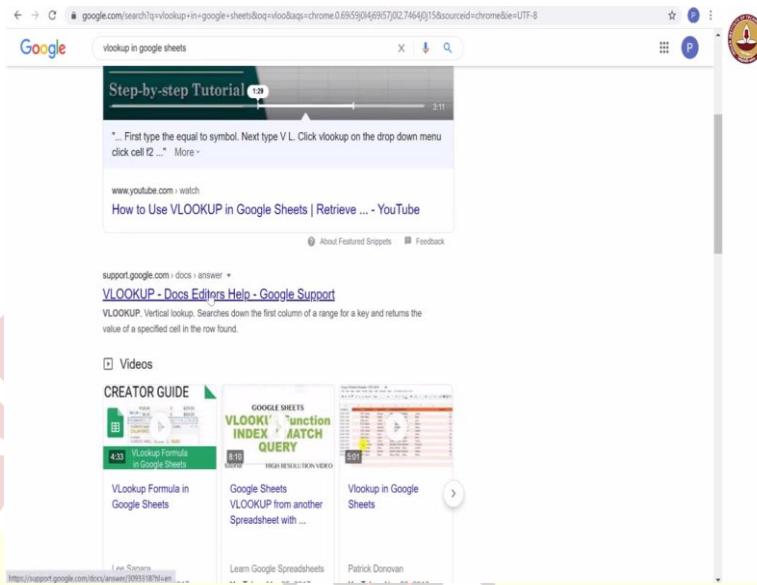
Statistics for Data Science
Professor. Usha Mohan
Prathyush P (Support Team)
Department of Management Studies
Indian Institute of Technology, Madras
Tutorial – 4
VLOOKUP in Google Sheets

(Refer Slide Time: 00:14)



Hello Statistics students. In this tutorial, we will look up a particular very useful function in Google Sheets. And that is called VLOOKUP, which stands for vertical lookup. So, let us see what it does. And for that, let us Google VLOOKUP.

(Refer Slide Time: 00:37)



And here, let us go to the support.google.com, the official Google support for docs and let us see what they have to say.

(Refer Slide Time: 00:48)

VLOOKUP
Vertical lookup. Searches down the first column of a range for a key and returns the value of a specified cell in the row found.

Sample Usage
`VLOOKUP(10003, A2:B26, 2, FALSE)`

Syntax

```
VLOOKUP(search_key, range, index, [is_sorted])
• search_key - The value to search for. For example, 42, "Cats", or I24.
• range - The range to consider for the search. The first column in the range is searched for the key specified in search_key.
• index - The column index of the value to be returned, where the first column in range is numbered 1.
• If index is not between 1 and the number of columns in range, #VALUE! is returned.
• is_sorted - [TRUE by default] - Indicates whether the column to be searched (the first column of the specified range) is sorted. FALSE is recommended in most cases.
• It's recommended to set is_sorted to FALSE. If set to FALSE, an exact match is returned. If there are multiple matching values, the content of the cell corresponding to the first value found is returned, and #N/A is returned if no such value is found.
```

So as we discussed, VLOOKUP stands for Vertical Lookup. And what does it do, it searches down the first column of a range. So, we give a range of data in the spreadsheet. And it searches through the first column of this range for a specific key, and then returns the value of a specified cell in the row found. So, what we mean by this is something like this.

(Refer Slide Time: 01:21)

A screenshot of a Google Sheets spreadsheet titled 'Scores'. The data is organized into columns labeled A through P. Row 1 contains column headers: Card, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14. Rows 2 through 9 contain student information: Name, Gender, D.O.B., Place, and various subject scores (Math, Physics, Chem) along with a Total score. Row 10 is blank.

Card	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Student	huvanesh	Harish	Shashank	Rida	Ritika	Akshaya	Sameer	Aditya	Surya	Clarence	Kavya	Rahul	Srinidhi	Gopi	Sophia
Gender	M	M	M	F	F	F	M	M	M	M	F	M	F	M	F
D.O.B.	7 Nov	3 Jun	4 Jan	5 May	17 Nov	8 Feb	23 Mar	15 Mar	28 Feb	6 Dec	12 Jan	30 Apr	14 Jan	6 May	23 July
Place	Erode	Salem	Chennai	Chennai	Madurai	Chennai	Ambar	Vellore	Bengaluru	Bengaluru	Chennai	Bengaluru	Chennai	Madurai	Trichy
Math	68	62	57	42	87	71	81	84	74	63	64	97	52	65	89
Physics	64	45	54	53	64	92	82	92	64	88	72	92	64	73	62
Chem	78	91	77	78	89	84	87	76	51	73	68	92	71	89	93
Total	210	198	188	173	240	247	250	252	189	224	204	281	187	227	244

If you recollect, this is the scores data in Computational thinking course. So, each of these is a particular Card, each column has a particular Card, except the first one. The first one is just the names of the fields. So we have 29 as the last number, but we have beginning from 0 therefore, we have 30 cards overall. Now, we know that this is the Card number, this is the Student's name, this is Gender, this is Date of Birth, so on, so forth. Now the way to use VLOOKUP is let us assume we want the place off card number 7.

(Refer Slide Time: 02:09)

Card	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Student	huvanesh	Harish	Shashank	Rida	Ritika	Akshaya	Sameer	Aditya	Surya	Clarence	Kavya	Rahul	Srinidhi	Gopi	Sophia
Gender	M	M	M	F	F	F	M	M	M	M	F	M	F	M	F
D.O.B.	7 Nov	3 Jun	4 Jan	5 May	17 Nov	8 Feb	23 Mar	15 Mar	28 Feb	6 Dec	12 Jan	30 Apr	14 Jan	6 May	23 July
Place	Erode	Salem	Chennai	Chennai	Madurai	Chennai	Ambar	Vellore	Bengaluru	Bengaluru	Chennai	Bengaluru	Chennai	Madurai	Trichy
Math	68	62	57	42	87	71	81	84	74	63	64	97	52	65	89
Physics	64	45	54	53	64	92	82	92	64	88	72	92	64	73	62
Chem	78	91	77	78	89	84	87	76	51	73	68	92	71	89	93
Total	210	198	188	173	240	247	250	252	189	224	204	281	187	227	244

So, that would be Vellore. So here, we are able to see directly, but suppose we cannot really go through the whole list. So, we would do this is equal to VLOOKUP. And the first parameter we are supposed to give is the search key. So, what is the value in the first column that the function should look for, the value to search for?

(Refer Slide Time: 02:42)

	A	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	Card	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
2	Student	Goutami	Tauseef	Arshad	Abirami	Vetrivel	Kalyan	Monika	Priya	Deepika	Siddharth	Geeta	JK	Jagan	Nisha	Naveen
3	Gender	F	M	M	F	M	M	F	F	F	M	F	M	M	F	M
4	D.O.B.	22 Sep	30 Dec	14 Dec	9 Oct	30 Aug	17 Sep	15 Mar	17 Jul	13 May	26 Dec	16 May	22 Jul	4 Mar	10 Sep	13 Oct
5	Place	Theni	Trichy	Chennai	Erode	Trichy	Vellore	Bengaluru	Nagercoil	Bengaluru	Madurai	Chennai	Chennai	Madurai	Madurai	Vellore
6	Math	76	87	62	72	56	93	78	62	97	44	87	74	81	74	72
7	Physics	58	86	81	92	78	68	69	62	91	72	75	71	76	83	66
8	Chem	90	43	67	97	62	91	74	57	88	58	92	82	52	83	81
9	Total	224	216	210	261	196	252	221	181	276	174	254	227	209	240	219
10																
11																
12																
13																
14																
15																
16																

So, here the value to search for is Place so this is the search key, and then the second parameter, we have the range of data, so what is the entire range of data? So, for this, we can select the entire data we have and that would be the range and then we go to the next parameter, which is index, what does this index mean the index is the column index of the value to be returned.

(Refer Slide Time: 03:23)

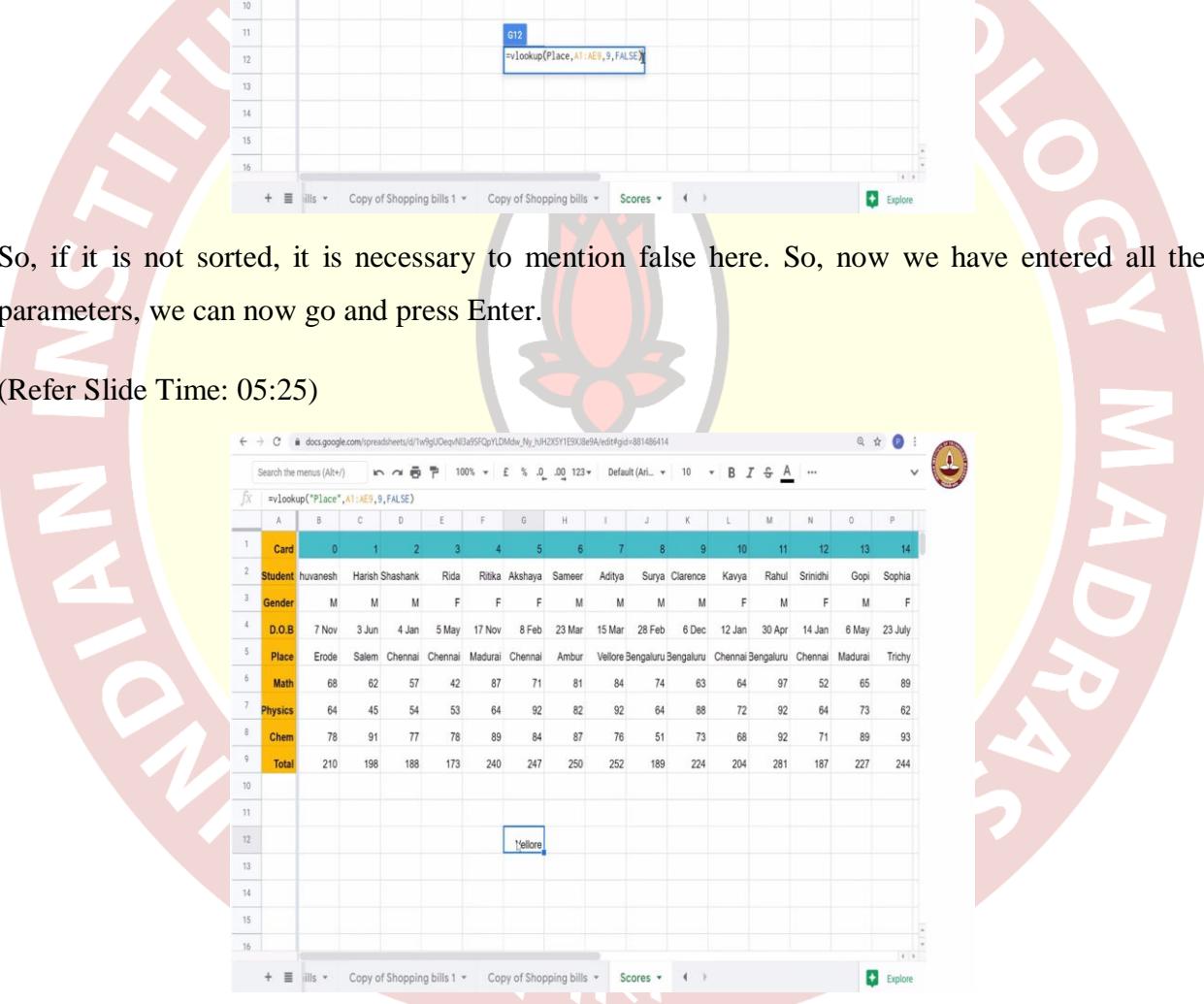
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Card	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	Student	huvanesh	Harish	Shashank	Rida	Ritika	Akshaya	Sameer	Aditya	Surya	Clarence	Kavya	Rahul	Srinidhi	Gopi	Sophia
3	Gender	M	M	M	F	F	F	M	M	M	M	M	F	M	F	M
4	D.O.B.	7 Nov	3 Jun	4 Jan	5 May	17 Nov	8 Feb	23 Mar	15 Mar	28 Feb	6 Dec	12 Jan	30 Apr	14 Jan	6 May	23 July
5	Place	Erode	Salem	Chennai	Chennai	Madurai	Chennai	Ambur	Vellore	Bengaluru	Bengaluru	Chennai	Bengaluru	Chennai	Madurai	Trichy
6	Math	68	62	57	42	87	71	81	84	74	63	64	97	52	65	89
7	Physics	64	45	54	53	64	92	82	92	64	88	72	92	64	73	62
8	Chem	78	91	77	78	89	84	87	76	51	73	68	92	71	89	93
9	Total	210	198	188	173	240	247	250	252	189	224	204	281	187	227	244
10																
11																
12									G12	=vlookup(Place, A1:I, I:I)						
13																
14																
15																
16																

So, we are trying to find specifically the place of this particular card and here now, the index for that would be our A column is the first column. So, this is second, this is third, this is four, this is fifth, the sixth, seventh, eighth and ninth, so 9 is the index column. So, what are we doing we are basically trying to find the key place in the first column, and then look at what is the value corresponding to that row in the ninth column, this is what we are trying to do.

And then the final key is, is_sorted. So, is_sorted is a binary variable which is set to true by default. So, it can take only two values either it is sorted, so it is true. And it is or it is not sorted and it is or it is not sorted and thus it is false. So, what is sorted here is this indicates whether the column to be searched, which is the first column of the specified range is sorted.

So, here we are basically asking if these values are sorted. What do we mean by sorted are they arranged in an order in ascending or descending order? Here that is not true. They are not arranged in any order. So, we give False if they were arranged in an order we could give true. And if we just do not give that parameter at all the function assumes that true is default the value. So, this is true by default.

(Refer Slide Time: 05:17)

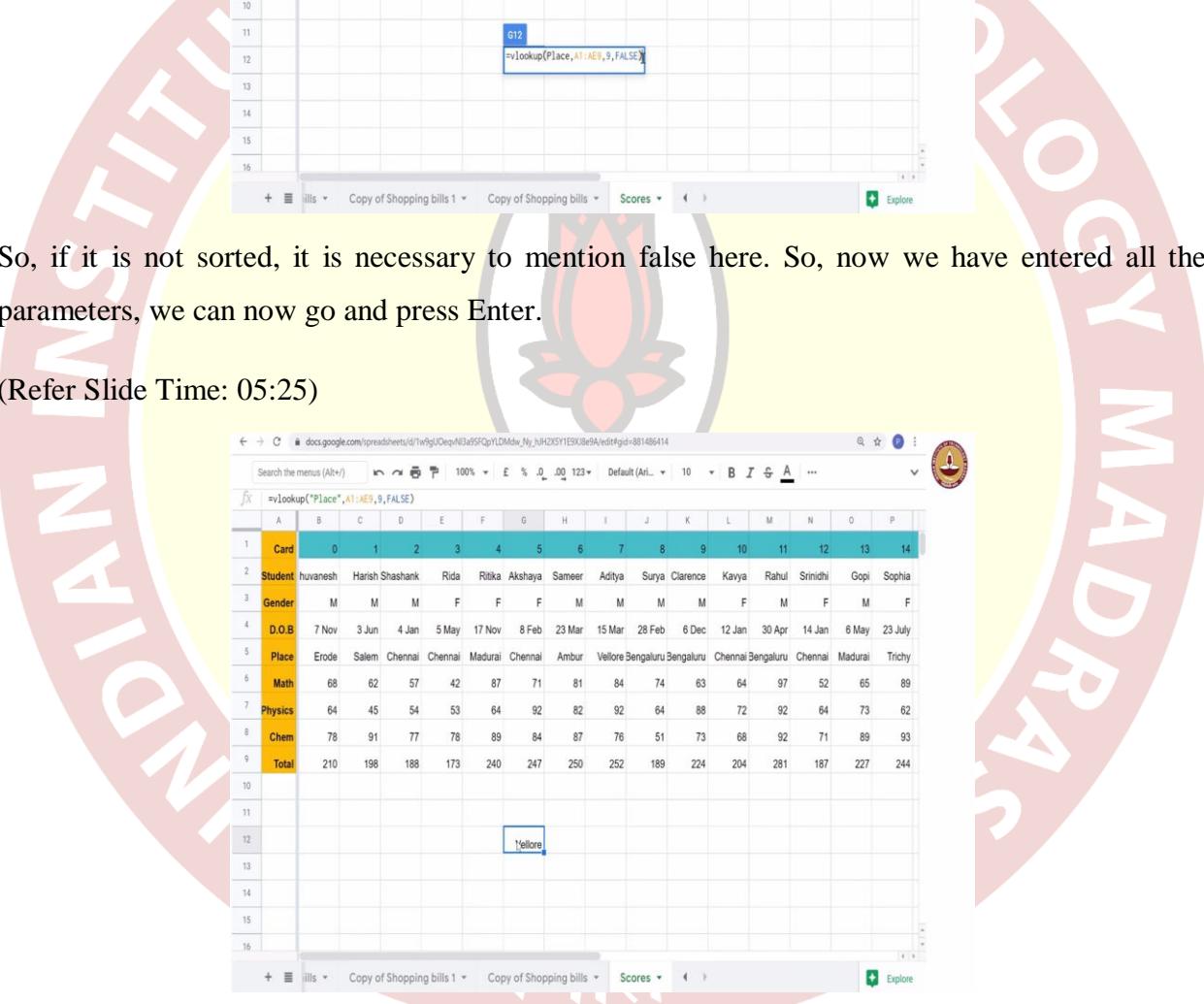


A screenshot of a Google Sheets document titled "Copy of Shopping bills 1". The sheet contains a table with student information. The formula `=vlookup("Place",A1:A14,9,FALSE)` is entered in cell G12. The table has columns for Card (0-14), Student (huvanesh, Harish, Shashank, Rida, Ritika, Akshaya, Sameer, Aditya, Surya, Clarence, Kavya, Rahul, Srinidhi, Gopi, Sophia), Gender (M/F), D.O.B (7 Nov, 3 Jun, 4 Jan, 5 May, 17 Nov, 8 Feb, 23 Mar, 15 Mar, 28 Feb, 6 Dec, 12 Jan, 30 Apr, 14 Jan, 6 May, 23 July), Place (Erode, Salem, Chennai, Madurai, Vellore, Bengaluru, Bengaluru, Chennai, Bengaluru, Chennai, Madurai, Trichy), and Math/Physics/Chem/Total scores.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Card	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	Student	huvanesh	Harish	Shashank	Rida	Ritika	Akshaya	Sameer	Aditya	Surya	Clarence	Kavya	Rahul	Srinidhi	Gopi	Sophia
3	Gender	M	M	M	F	F	F	M	M	M	M	F	M	F	M	F
4	D.O.B	7 Nov	3 Jun	4 Jan	5 May	17 Nov	8 Feb	23 Mar	15 Mar	28 Feb	6 Dec	12 Jan	30 Apr	14 Jan	6 May	23 July
5	Place	Erode	Salem	Chennai	Chennai	Madurai	Chennai	Ambar	Vellore	Bengaluru	Bengaluru	Chennai	Bengaluru	Chennai	Madurai	Trichy
6	Math	68	62	57	42	87	71	81	84	74	63	64	97	52	65	89
7	Physics	64	45	54	53	64	92	82	92	64	88	72	92	64	73	62
8	Chem	78	91	77	78	89	84	87	76	51	73	68	92	71	89	93
9	Total	210	198	188	173	240	247	250	252	189	224	204	281	187	227	244
10																
11																
12																
13																
14																
15																
16																

So, if it is not sorted, it is necessary to mention false here. So, now we have entered all the parameters, we can now go and press Enter.

(Refer Slide Time: 05:25)



A screenshot of a Google Sheets document titled "Copy of Shopping bills 1". The sheet contains a table with student information. The formula `=vlookup("Place",A1:A14,9,FALSE)` is entered in cell G12. The table has columns for Card (0-14), Student (huvanesh, Harish, Shashank, Rida, Ritika, Akshaya, Sameer, Aditya, Surya, Clarence, Kavya, Rahul, Srinidhi, Gopi, Sophia), Gender (M/F), D.O.B (7 Nov, 3 Jun, 4 Jan, 5 May, 17 Nov, 8 Feb, 23 Mar, 15 Mar, 28 Feb, 6 Dec, 12 Jan, 30 Apr, 14 Jan, 6 May, 23 July), Place (Erode, Salem, Chennai, Madurai, Vellore, Bengaluru, Bengaluru, Chennai, Bengaluru, Chennai, Madurai, Trichy), and Math/Physics/Chem/Total scores.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Card	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	Student	huvanesh	Harish	Shashank	Rida	Ritika	Akshaya	Sameer	Aditya	Surya	Clarence	Kavya	Rahul	Srinidhi	Gopi	Sophia
3	Gender	M	M	M	F	F	F	M	M	M	M	F	M	F	M	F
4	D.O.B	7 Nov	3 Jun	4 Jan	5 May	17 Nov	8 Feb	23 Mar	15 Mar	28 Feb	6 Dec	12 Jan	30 Apr	14 Jan	6 May	23 July
5	Place	Erode	Salem	Chennai	Chennai	Madurai	Chennai	Ambar	Vellore	Bengaluru	Bengaluru	Chennai	Bengaluru	Chennai	Madurai	Trichy
6	Math	68	62	57	42	87	71	81	84	74	63	64	97	52	65	89
7	Physics	64	45	54	53	64	92	82	92	64	88	72	92	64	73	62
8	Chem	78	91	77	78	89	84	87	76	51	73	68	92	71	89	93
9	Total	210	198	188	173	240	247	250	252	189	224	204	281	187	227	244
10																
11																
12																
13																
14																
15																
16																

And we have an error here the error is because Place is a text, it is a string value, so, we should put it in these quotation marks, and now we press Enter we get Vellore.

(Refer Slide Time: 05:48)

A screenshot of a Google Sheets document titled 'Copy of Shopping bills 1'. The sheet contains a table with student information and scores. A formula is being typed into cell A12: `=vlookup("0.0.B",A1:A14,9, FALSE)`. The table includes columns for Card (highlighted in yellow), Student (huvanesh, Harish, Shashank, Rida, Ritika, Akshaya, Sameer, Aditya, Surya, Clarence, Kavya, Rahul, Srinidhi, Gopi, Sophia), Gender (M, M, M, F, F, F, M, M, M, M, M, F, M, F, M, F), D.O.B (7 Nov, 3 Jun, 4 Jan, 5 May, 17 Nov, 8 Feb, 23 Mar, 15 Mar, 28 Feb, 6 Dec, 12 Jan, 30 Apr, 14 Jan, 6 May, 23 July), Place (Erode, Salem, Chennai, Chennai, Madurai, Chennai, Ambur, Vellore, Bengaluru, Bengaluru, Chennai, Bengaluru, Chennai, Madurai, Trichy), and subjects (Math, Physics, Chem) with their respective marks. The total column shows the sum of marks for each student.

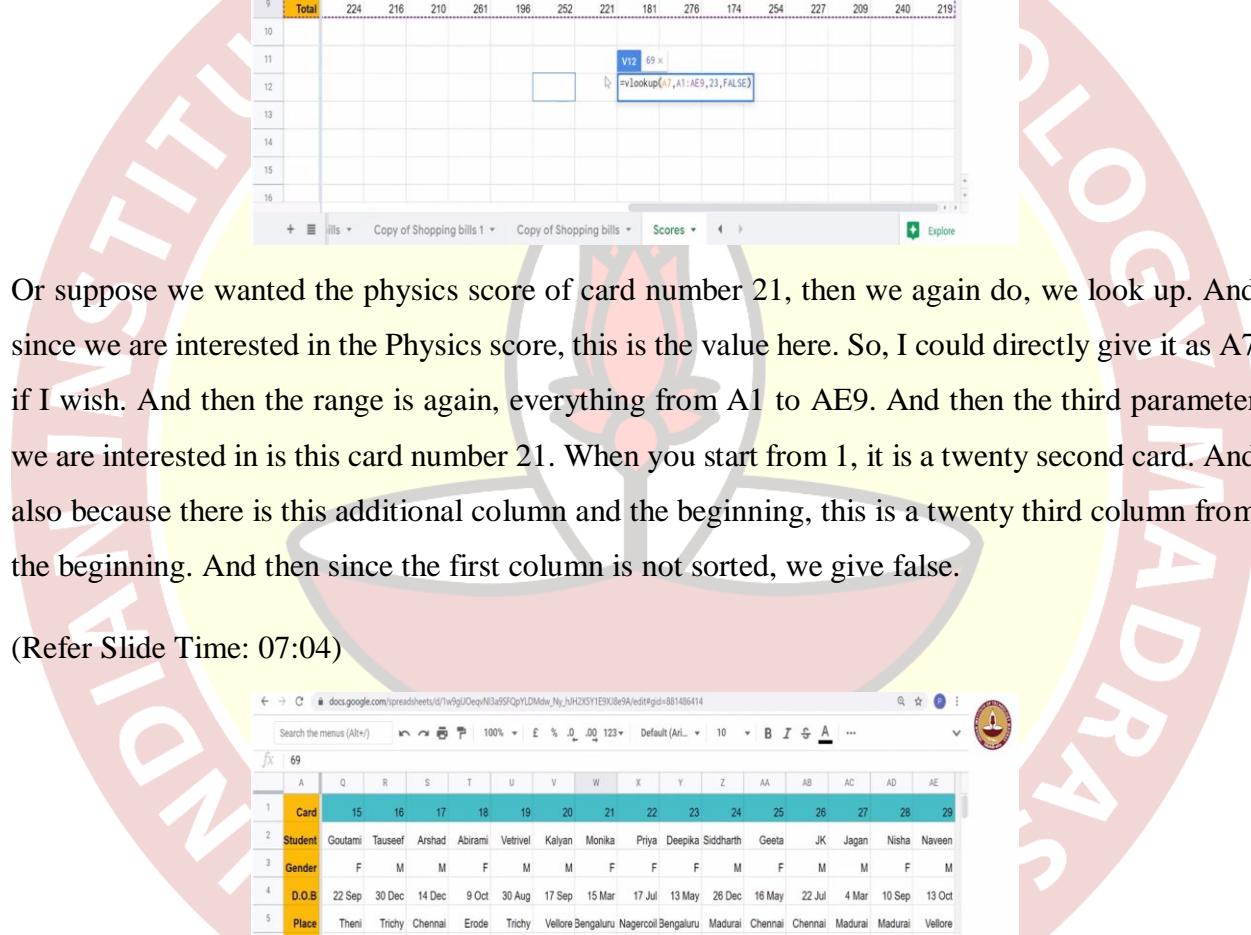
So, if you want to try it again, instead of Place, let us give Date of Birth.

(Refer Slide Time: 05:55)

A screenshot of a Google Sheets document titled 'Copy of Shopping bills 1'. The sheet contains a table with student information and scores. A formula is being typed into cell A12: `=vlookup("0.0.B",A1:A14,9, FALSE)`. The table includes columns for Card (highlighted in yellow), Student (huvanesh, Harish, Shashank, Rida, Ritika, Akshaya, Sameer, Aditya, Surya, Clarence, Kavya, Rahul, Srinidhi, Gopi, Sophia), Gender (M, M, M, F, F, F, M, M, M, M, M, F, M, F, M, F), D.O.B (7 Nov, 3 Jun, 4 Jan, 5 May, 17 Nov, 8 Feb, 23 Mar, 15 Mar, 28 Feb, 6 Dec, 12 Jan, 30 Apr, 14 Jan, 6 May, 23 July), Place (Erode, Salem, Chennai, Chennai, Madurai, Chennai, Ambur, Vellore, Bengaluru, Bengaluru, Chennai, Bengaluru, Chennai, Madurai, Trichy), and subjects (Math, Physics, Chem) with their respective marks. The total column shows the sum of marks for each student. The date 15 Mar is highlighted in the formula bar.

And now we get 15 March, which is this.

(Refer Slide Time: 06:01)

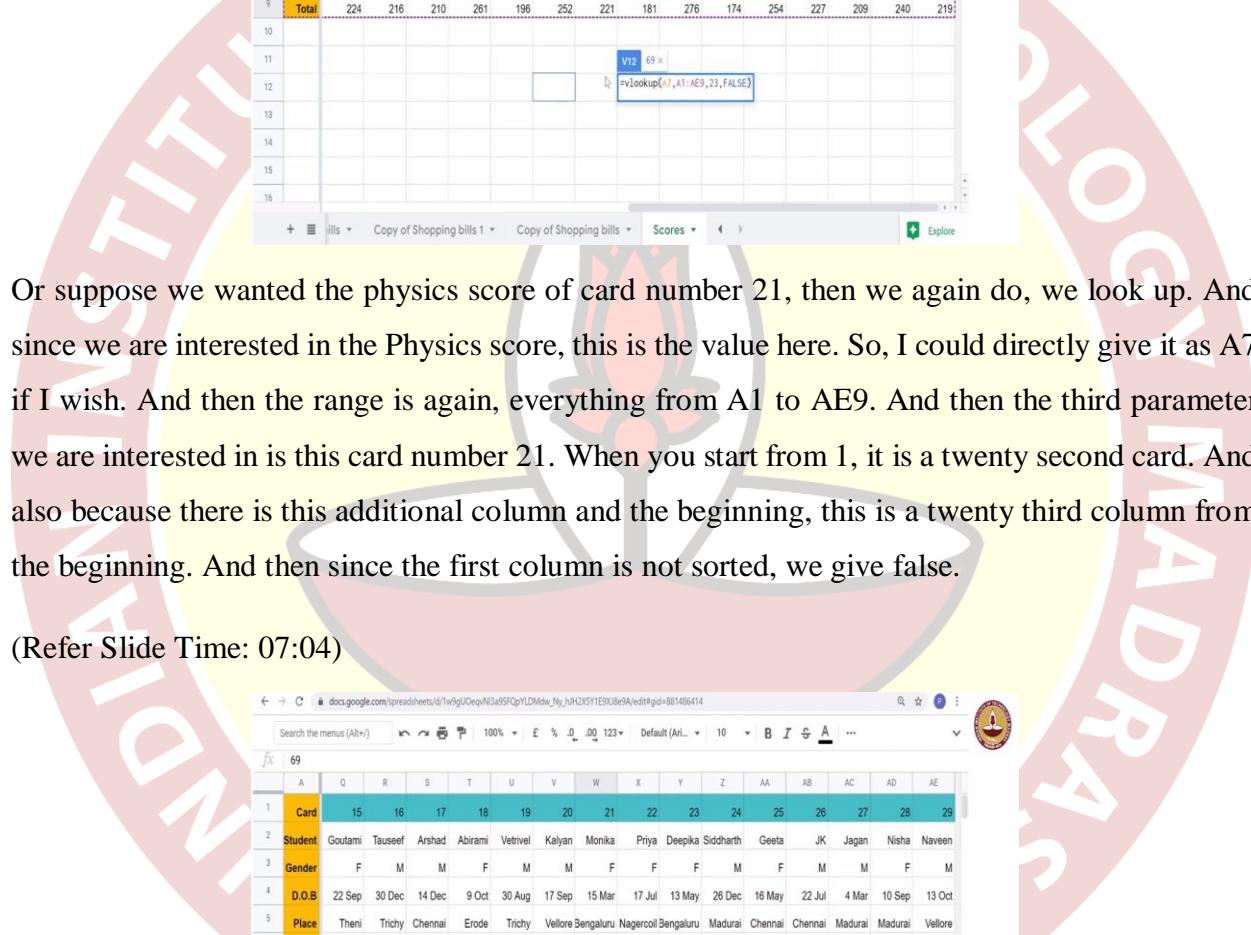


A screenshot of a Google Sheets spreadsheet titled 'Scores'. The spreadsheet contains data for 29 students across various subjects. The columns are labeled from A to AE. Row 1 contains subject names like Card, Student, Gender, D.O.B., Place, Math, Physics, Chem, and Total. Rows 2 through 9 list student details such as name, gender, DOB, place, and scores for Math, Physics, and Chem. Row 10 is blank. Row 11 contains the formula '=vlookup(A7,A1:AE9,23,FALSE)' in cell V11. Row 12 contains the result '69'. The formula is also highlighted with a blue border.

	A	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	Card	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
2	Student	Goutami	Tauseef	Arshad	Abirami	Vetrivel	Kalyan	Monika	Priya	Deepika	Siddharth	Geeta	JK	Jagan	Nisha	Naveen
3	Gender	F	M	M	F	M	M	F	F	F	M	F	M	M	F	M
4	D.O.B.	22 Sep	30 Dec	14 Dec	9 Oct	30 Aug	17 Sep	15 Mar	17 Jul	13 May	26 Dec	16 May	22 Jul	4 Mar	10 Sep	13 Oct
5	Place	Theni	Trichy	Chennai	Erode	Trichy	Vellore	Bengaluru	Nagercoil	Bengaluru	Madurai	Chennai	Chennai	Madurai	Madurai	Vellore
6	Math	76	87	62	72	56	93	78	62	97	44	87	74	81	74	72
7	Physics	58	86	81	92	78	68	69	62	91	72	75	71	76	83	66
8	Chem	90	43	67	97	62	91	74	57	88	58	92	82	52	83	81
9	Total	224	216	210	261	196	252	221	181	276	174	254	227	209	240	219
10																
11																
12																
13																
14																
15																
16																

Or suppose we wanted the physics score of card number 21, then we again do, we look up. And since we are interested in the Physics score, this is the value here. So, I could directly give it as A7 if I wish. And then the range is again, everything from A1 to AE9. And then the third parameter we are interested in is this card number 21. When you start from 1, it is a twenty second card. And also because there is this additional column and the beginning, this is a twenty third column from the beginning. And then since the first column is not sorted, we give false.

(Refer Slide Time: 07:04)



A screenshot of a Google Sheets spreadsheet titled 'Scores'. The spreadsheet contains data for 29 students across various subjects. The columns are labeled from A to AE. Row 1 contains subject names like Card, Student, Gender, D.O.B., Place, Math, Physics, Chem, and Total. Rows 2 through 9 list student details such as name, gender, DOB, place, and scores for Math, Physics, and Chem. Row 10 is blank. Row 11 contains the formula '=vlookup(A7,A1:AE9,23,FALSE)' in cell V11. Row 12 contains the result '69'. The formula is also highlighted with a blue border.

	A	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	Card	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
2	Student	Goutami	Tauseef	Arshad	Abirami	Vetrivel	Kalyan	Monika	Priya	Deepika	Siddharth	Geeta	JK	Jagan	Nisha	Naveen
3	Gender	F	M	M	F	M	M	F	F	F	M	F	M	M	F	M
4	D.O.B.	22 Sep	30 Dec	14 Dec	9 Oct	30 Aug	17 Sep	15 Mar	17 Jul	13 May	26 Dec	16 May	22 Jul	4 Mar	10 Sep	13 Oct
5	Place	Theni	Trichy	Chennai	Erode	Trichy	Vellore	Bengaluru	Nagercoil	Bengaluru	Madurai	Chennai	Chennai	Madurai	Madurai	Vellore
6	Math	76	87	62	72	56	93	78	62	97	44	87	74	81	74	72
7	Physics	58	86	81	92	78	68	69	62	91	72	75	71	76	83	66
8	Chem	90	43	67	97	62	91	74	57	88	58	92	82	52	83	81
9	Total	224	216	210	261	196	252	221	181	276	174	254	227	209	240	219
10																
11																
12																
13																
14																
15																
16																

And so that way we got the Physics score of this particular card. So, this is how we can use we VLOOKUP.

(Refer Slide Time: 07:18)

	A	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	Card	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
2	Student	Goutami	Tauseef	Arshad	Abirami	Retrivel	Kalyan	Monika	Priya	Deepika	Siddharth	Geeta	JK	Jagan	Nisha	Naveen
3	Gender	F	M	M	F	M	M	F	F	F	M	F	M	M	F	M
4	D.O.B	22 Sep	30 Dec	14 Dec	9 Oct	30 Aug	17 Sep	15 Mar	17 Jul	13 May	26 Dec	16 May	22 Jul	4 Mar	10 Sep	13 Oct
5	Place	Theni	Trichy	Chennai	Erode	Trichy	Vellore	Bengaluru	Nagercoil	Bengaluru	Madurai	Chennai	Chennai	Madurai	Madurai	Vellore
6	Math	76	87	62	72	56	93	78	62	97	44	87	74	81	74	72
7	Physics	58	86	81	92	78	68	69	62	91	72	75	71	76	83	66
8	Chem	90	43	67	97	62	91	74	57	88	58	92	82	52	83	81
9	Total	224	216	210	261	196	252	221	181	276	174	254	227	209	240	219
10																
11																
12																
13																
14																
15																
16																

However, in the formula, when we use A7 here, we know that the seventh row is what we are looking for. In general, we would not in general, you can have thousands of rows, and you do not know which row matches your requirement. So, it is usually best to type in the actual value over there in the first parameter. Now let us go back to our

(Refer Slide Time: 07:44)

Examples		
In this example, VLOOKUP searches down the first column for a student ID and returns the corresponding grade		
VLOOKUP : Student Grade		
Student ID	Grade	Formula
N444	90	=VLOOKUP(A9,\$A\$2:\$B\$5, 2, FALSE)
N333	100	=VLOOKUP(A10,\$A\$2:\$B\$5, 2, FALSE)
N222	85	=VLOOKUP(A11,\$A\$2:\$B\$5, 2, FALSE)
N111	80	=VLOOKUP(A12,\$A\$2:\$B\$5, 2, FALSE)
Student Grade		
Make a copy		
In this example, VLOOKUP searches down the first column for the income using approximate match (<code>is_sorted</code> is set to TRUE) and returns the corresponding tax rate		
VLOOKUP : Income Tax Rate		
Lower bound Upper bound Try Data		

Doc editor's help. And there are a couple of examples here that we should be looking into these are of some value. In this example, VLOOKUP searches down the first column for a student ID and returns a corresponding grade. So here, there are these student IDs, these 4 and their respective grades. And now here the student ID is entered, and their grade is being brought up by the VLOOKUP formula. As you can see, the VLOOKUP formula is being used for A9, A10, A11 and A12 respectively, which are these ID numbers. So, this is kind of like what we have already seen.

(Refer Slide Time: 08:29)

The screenshot shows a Google Docs Editors Help page with the URL support.google.com/docs/answer/3093318?hl=en. The page displays a VLOOKUP example. At the top, there is a table with columns 'Lower bound', 'Upper bound', and 'Tax Rate'. Below it is another table with columns 'Title', 'Income', 'Tax Rate', and 'Formula'. The 'Formula' column shows how VLOOKUP is used to find tax rates based on income levels. A note at the bottom states: "In this example, VLOOKUP searches down the first column for the income using approximate match (is_sorted is set to TRUE) and returns the corresponding tax rate".

The more interesting example comes here. Now let us make a copy of this and see what is interesting about it.

(Refer Slide Time: 08:37)

The screenshot shows a Google Sheets interface with a circular watermark in the background reading "INDIAN INSTITUTE OF LOGIC MADRAS". At the top, there's a "Copy document" dialog box with a "Make a copy" button. Below it, the main spreadsheet area displays two tables. The first table is a tax rate lookup table:

	A	B	C	D	E	F	G
1	Lower bound	Upper bound	Tax Rate				
2	\$0	\$2,999.99	1%				
3	\$3,000	\$5,999.99	2%				
4	\$6,000	\$9,999.99	3%				
5	\$10,000	\$13,999.99	4%				
6	\$14,000		5%				

The second table lists job titles, incomes, and their corresponding formulas:

	Title	Income	Tax Rate	Formula
10	Associate	\$3,300	2%	=VLOOKUP(B10, \$A\$2:\$C\$6, 3, TRUE)
11	Vice President	\$6,500	3%	=VLOOKUP(B11, \$A\$2:\$C\$6, 3, TRUE)
12	Director	\$9,000	3%	=VLOOKUP(B12, \$A\$2:\$C\$6, 3, TRUE)
13	Managing Director	\$11,000	4%	=VLOOKUP(B13, \$A\$2:\$C\$6, 3, TRUE)

Here in this example, we are looking at Tax rates, the Tax percentages for incomes in these ranges 0 to 2,999.9. So, less than 3000 has a tax rate of 1% and 3000 to 5999.99 which is less than 6000 but greater than or equal to 3000 is 2% and then 6000 to less than 10,000 is 3% and then 10,000 to less than 14,000 is 4%. And lastly, anything greater than 14,000 or equal to 14,000 is 5%. So, this is the data.

Now here we are making a table with the titles of Jobs. So, there is associate as vice president as director and as managing director, and we are also putting in their respective incomes. Now, as you can see, these values 3300, 6500, 9000 and 11,000 do not actually exist in any of these values

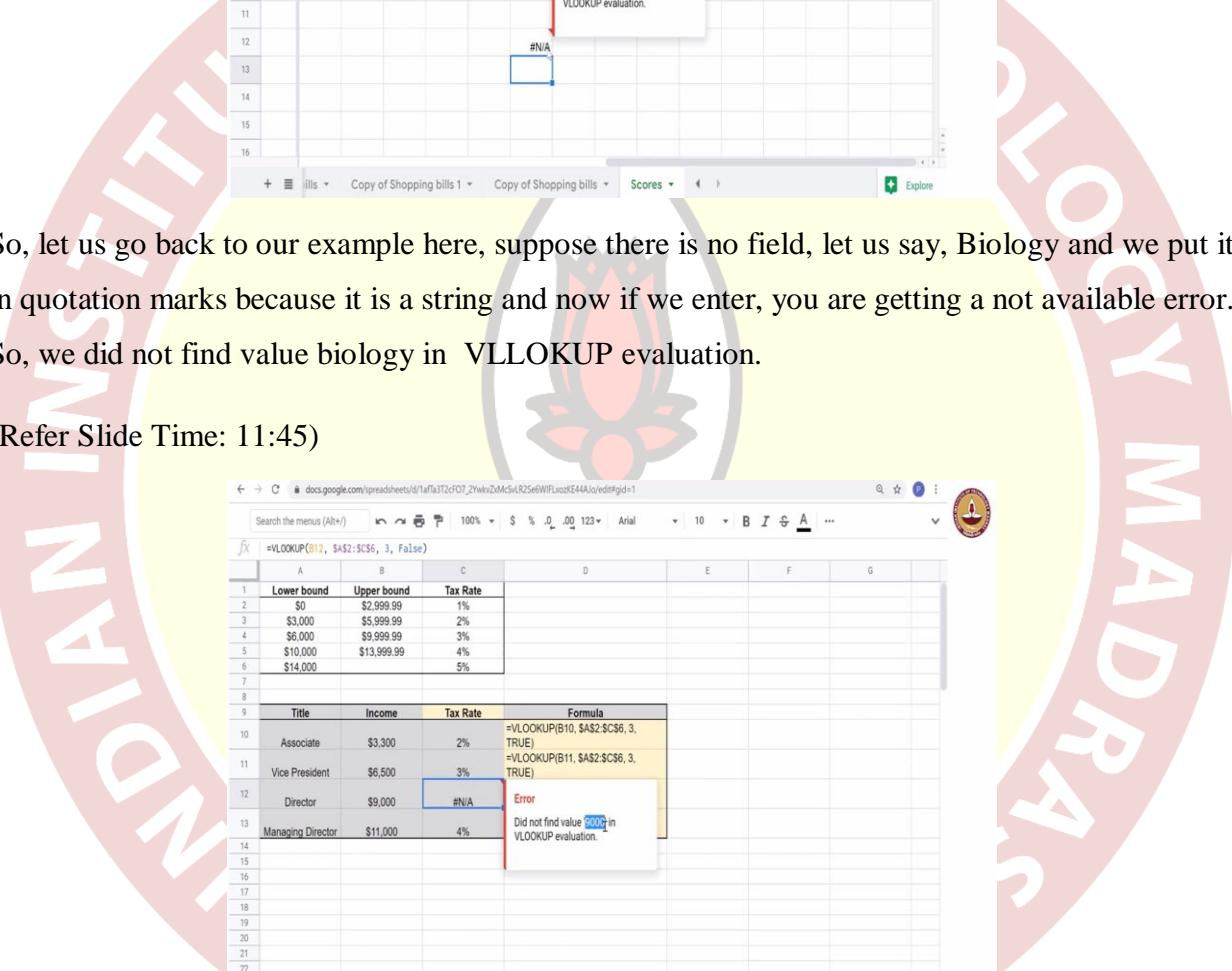
here. So, what VLOOKUP does is we now also know that this is a sorted list, so we give the true as the last parameter. True for the last parameter it is sorted. And then what VLOOKUP does is if it is a sorted column, it will approximately find what you are looking for.

So, 3300 does not actually exist in this particular data in the first column, but 3300 is between 3000 and 6000. So, it will pick up the 3000 category and put in the 2% that is expected. Likewise, 6500 is between 6000 and 10,000. So, it will pick up the tax rate for the 6000 to 10,000 slab and then 9000 also the same thing happens it is between 6000 and 10,000.

So, the 6000 tax rate is picked up and all of this is happening this approximation is happening because we set the last parameters true because the first column is sorted, it can do this if it is not sorted, if you give false that case, VLOOKUP will look exactly for the value you are giving. So, 3300 is what it will exactly look for and if it does not exist, what will happen then.

(Refer Slide Time: 11:16)

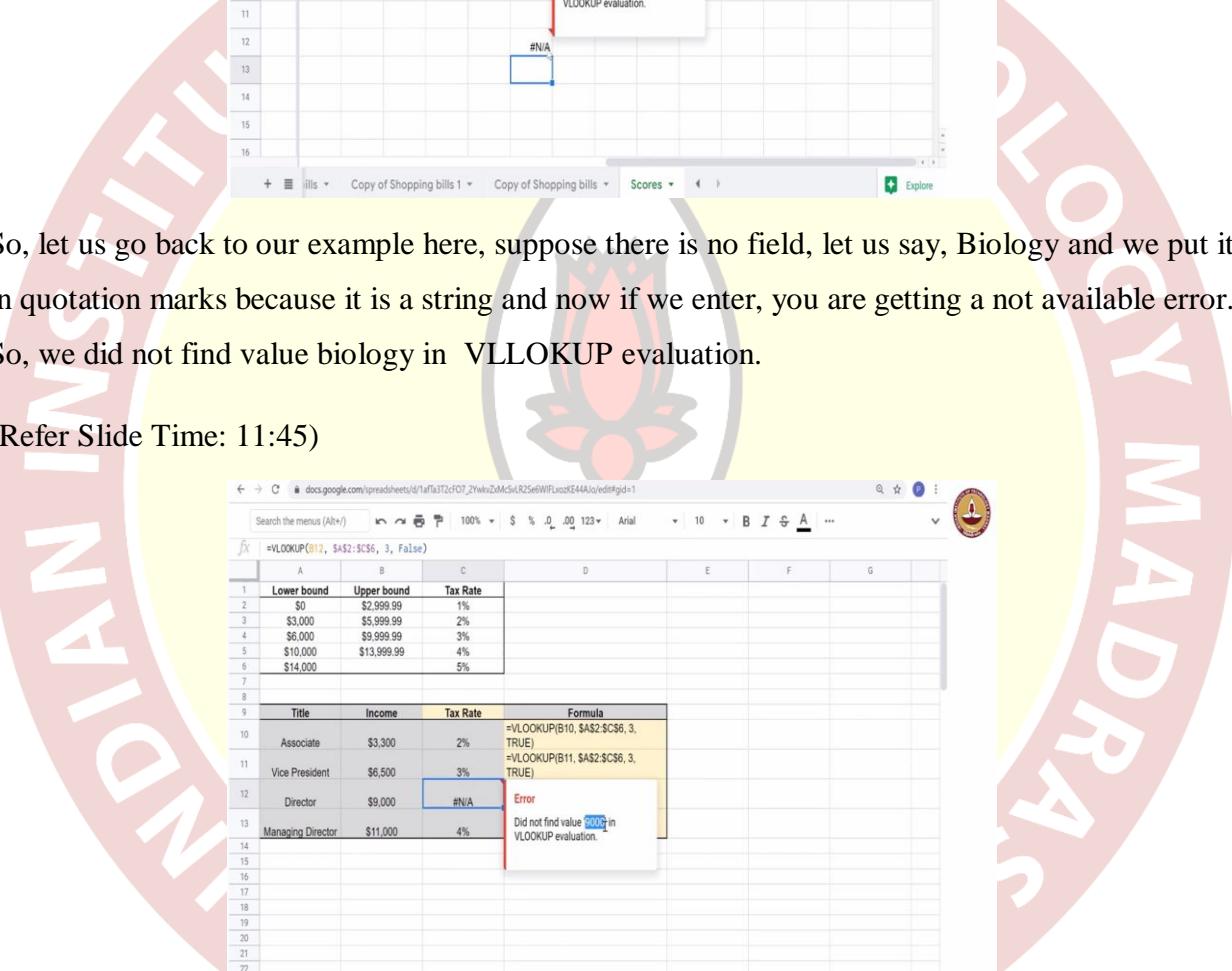
	A	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	Card	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
2	Student	Goutami	Tauseef	Arshad	Abirami	Vetrivel	Kalyan	Monika	Priya	Deepika	Siddharth	Geeta	JK	Jagan	Nisha	Naveen
3	Gender	F	M	M	F	M	M	F	F	F	M	F	M	M	F	M
4	D.O.B.	22 Sep	30 Dec	14 Dec	9 Oct	30 Aug	17 Sep	15 Mar	17 Jul	13 May	26 Dec	16 May	22 Jul	4 Mar	10 Sep	13 Oct
5	Place	Theni	Trichy	Chennai	Erode	Trichy	Vellore	Bengaluru	Nagercoil	Bengaluru	Madurai	Chennai	Chennai	Madurai	Madurai	Vellore
6	Math	76	87	62	72	56	93	78	62	97	44	87	74	81	74	72
7	Physics	58	86	81	92	78	68	69	62	91	72	75	71	76	83	66
8	Chem	90	43	67	97	62	91	74	57	88	58	92	82	52	83	81
9	Total	224	216	210	261	196	252	221	181	276	174	254	227	209	240	219
10																
11																
12																=vlookup("Biology",A1:A69,23,FALSE)
13																
14																
15																
16																



A screenshot of a Google Sheets spreadsheet titled 'Copy of Shopping bills 1'. The data includes columns for Card Number (A15-A29), Student Name (B15-B29), Gender (C15-C29), D.O.B (D15-D29), Place (E15-E29), and various subjects like Math, Physics, and Chem with their respective scores (F15-F29). A formula in cell G15 is causing an error: =VLOOKUP(\$B15,\$A\$2:\$C\$6,3, False). A tooltip from the formula says: 'Did not find value "Biology" in VLOOKUP evaluation.' The cell contains '#N/A'.

So, let us go back to our example here, suppose there is no field, let us say, Biology and we put it in quotation marks because it is a string and now if we enter, you are getting a not available error. So, we did not find value biology in VLOOKUP evaluation.

(Refer Slide Time: 11:45)



A screenshot of a Google Sheets spreadsheet titled 'Income Tax Rate'. It shows a tax rate table with columns for Lower bound, Upper bound, and Tax Rate. A formula in cell D10 is causing an error: =VLOOKUP(B10,\$A\$2:\$C\$6,3, False). A tooltip from the formula says: 'Did not find value 9000 in VLOOKUP evaluation.' The cell contains '#N/A'.

So here also, in this example, if we were to change this formula, let us say we change this formula from true to false, then you get an error and says that it cannot find 9000 in

VLOOKUP evaluation.

(Refer Slide Time: 12:04)

A screenshot of a Google Sheets spreadsheet. The top row contains the formula `=VLOOKUP($B1, A2:C6, 3, TRUE)`. Below this is a table with columns 'Lower bound', 'Upper bound', and 'Tax Rate'. The data rows are:

Lower bound	Upper bound	Tax Rate
\$0	\$2,999.99	1%
\$3,000	\$5,999.99	2%
\$6,000	\$9,999.99	3%
\$10,000	\$13,999.99	4%
\$14,000		5%

Below this is another table with columns 'Title', 'Income', 'Tax Rate', and 'Formula'. The data rows are:

Title	Income	Tax Rate	Formula
Associate	\$3,300	2%	=VLOOKUP(B10, \$A\$2:\$C\$6, 3, TRUE)
Vice President	\$6,500	3%	=VLOOKUP(B11, \$A\$2:\$C\$6, 3, TRUE)
Director	\$9,000	3%	=VLOOKUP(B12, \$A\$2:\$C\$6, 3, TRUE)
Managing Director	\$11,000	4%	=VLOOKUP(B13, \$A\$2:\$C\$6, 3, TRUE)

This would not be the case if we just kept this is true. And obviously this only applies when your first column is actually sorted. In that case, it will go for the approximate value. And we are getting 3% as required.

(Refer Slide Time: 12:20)

A screenshot of a Google Sheets spreadsheet. The top row contains the formula `=VLOOKUP(A10, A2:C6, 3, FALSE)`. Below this is a table with columns 'Guest Name', 'Table Number', 'Dietary Restriction', and 'Sent Invitation'. The data rows are:

Guest Name	Table Number	Dietary Restriction	Sent Invitation
David	2	Vegetarian	No
Bob	3	None	No
david	1	None	Yes
Nancy	4	None	No
Mary	5	Vegetarian	Yes

Below this is another table with columns 'Guest Name', 'Dietary Restriction', and 'Formula'. The data rows are:

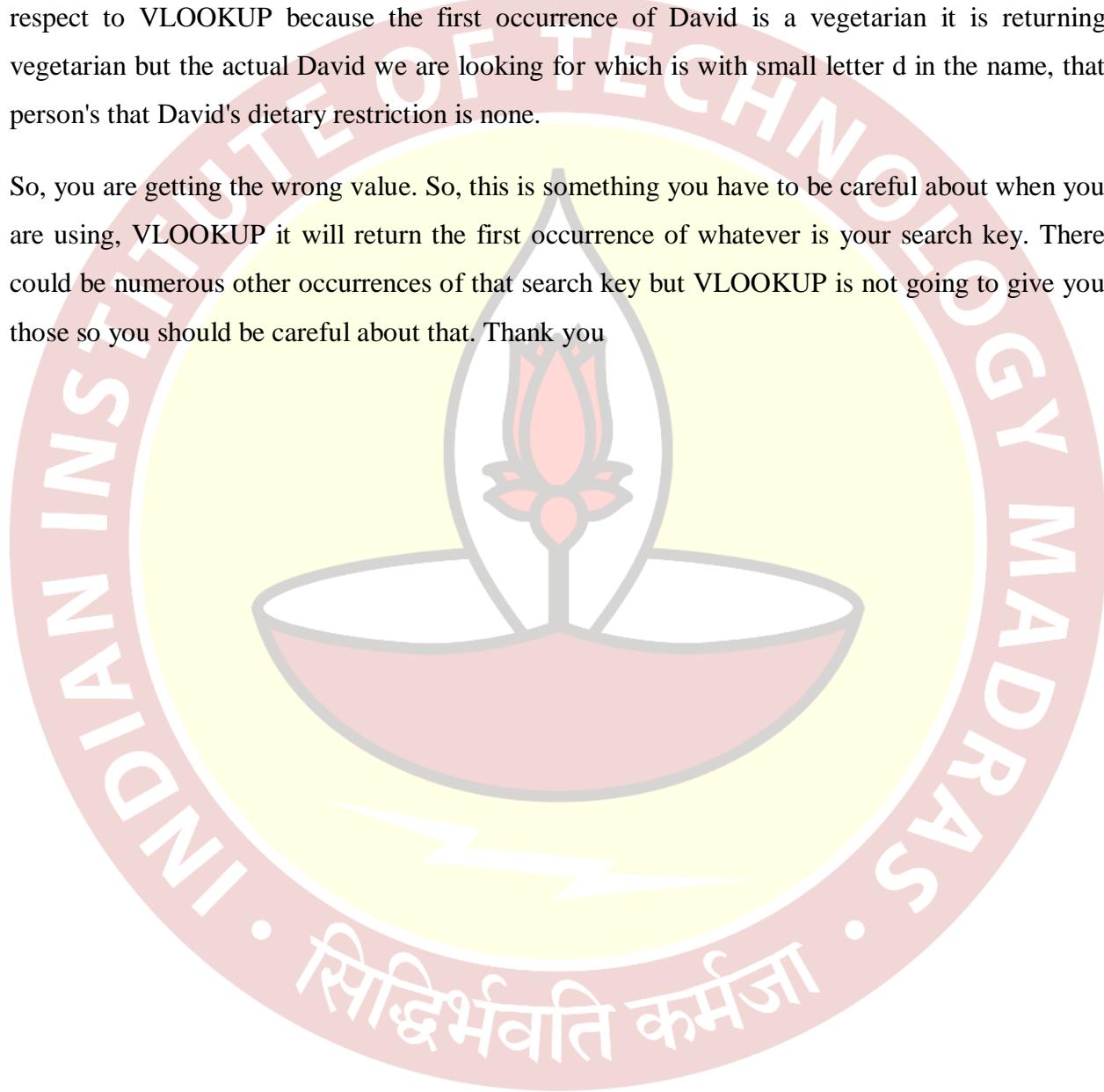
Guest Name	Dietary Restriction	Formula
david	Vegetarian	=VLOOKUP(A10, \$A\$2:\$C\$6, 3, FALSE)
Nancy	None	=VLOOKUP(A11, \$A\$2:\$C\$6, 3, FALSE)
Mary	Vegetarian	=VLOOKUP(A12, \$A\$2:\$C\$6, 3, FALSE)

In the last example that the doc editors help is providing. We have this table, where there is a guest name, and what table they must be sitting on and what kind of diet and also the invitation is sent to them on or not. Now they are looking of the dietary restriction for different people. So, there is

David, Nancy and Mary. And what diet restriction do they have, and for this they are doing the, VLOOKUP formula. The trouble here is there are two David's, there is David 1 here.

And there is also David 2 here. So, what VLOOKUP does is when you look for David in this list, it will match with the first occurrence of that name. So, this you have to be very careful with, with respect to VLOOKUP because the first occurrence of David is a vegetarian it is returning vegetarian but the actual David we are looking for which is with small letter d in the name, that person's that David's dietary restriction is none.

So, you are getting the wrong value. So, this is something you have to be careful about when you are using, VLOOKUP it will return the first occurrence of whatever is your search key. There could be numerous other occurrences of that search key but VLOOKUP is not going to give you those so you should be careful about that. Thank you



Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 3.1
Describing Numerical Data – Frequency Tables for numerical data

In today's module, so we are going to start with Describing Numerical Data using for a again for a single variable. Just a brief review of what we have done so far. So, if you look at what we have done so far, we started by understanding what is statistics, we broadly introduced to the 2 main branches of statistics which is descriptive statistics and inferential statistics that you will learn as a beginner.

Then, we introduce the concept of a sample and a population and then afterwards we also said; what is the difference between a sample and a population. Then, we went on to understand how data is collected. Now, once you collect the data we did not focus too much on how to collect data, but once you collect the data we said that, you can actually put down this data and identify the variables and the observations and set up what is called a data set.

Once that is done, then we also discussed about the types of data. Broadly, you can classify data as numerical and categorical and then we also understood the difference between cross sectional and time series data. Finally, we looked at the measurement scales; namely, nominal, ordinal, interval and ratio scales of measurements.

(Refer Slide Time: 01:47)

Statistics for Data Science -1

Review

1. What is statistics?

- ▶ Descriptive statistics, inferential statistics.
- ▶ Distinguish between a sample and a population.

2. Understand how data are collected.

- ▶ Identify variables and cases (observations) in a data set

3. Types of data-

- ▶ classify data as categorical(qualitative) or numerical(quantitative) data.
- ▶ Understand cross-sectional versus time-series data.
- ▶ Measurement scales-nominal, ordinal, interval and ratio.

4. Describing categorical data

- ▶ Creating frequency tables, understanding relative frequency
- ▶ Creating pie charts and bar charts
- ▶ Understanding violations
- ▶ Descriptive measures of Mode and Median

Then, we went on to understand how to describe categorical data. When we talk about describing categorical data, the key point we have noticed are we first looked at what are the how to create frequency tables. Now, when we create a frequency tables we introduced the concept of what we call relative frequency. After constructing frequency tables we looked at the graphical summaries of data.

For this we focused on 2 main graphical summaries; namely, the pie charts and the bar charts. We also looked at what could be the possible violations that could occur in terms of misrepresenting data and how to overcome the same.

Finally, we looked at 2 descriptive measures which are used for categorical data; the mode and the median, while mode is applicable only when you have while mode is applicable for both the nominal and ordinal data, median can be defined only when you have ordinal data. So, this is where we stand at this point of time.

(Refer Slide Time: 03:07)



Frequency tables
Organizing numerical data

Graphical summaries
Histograms
Stem-and-leaf diagram

Numerical summaries
Measures of central tendency
Measures of dispersion
Percentiles

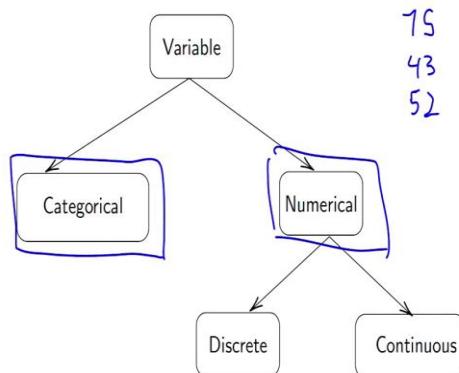


What we are going to do next is to understand how to summarize numerical data. The agenda is first we again will go through frequency tables, then we look at graphical summaries and then we will focus on the numerical summaries and measures of description.

(Refer Slide Time: 03:25)



Types of variables



Recall, when we talk about variables we broadly classified variables into the categorical variable and numerical variable. We have understood this portion. In the last module we understood how to describe categorical data. So, now, we are going to focus on how do

we describe or summarize numerical data. Again, when we look at numerical data they can be broadly classified into discrete and continuous. Now, what do we mean by discrete data? Discrete data as the name suggests for example, if I am looking at marks obtained by students and the marks obtained are 68, 75, 43; 52.

In a sense that these are discrete, but now when I am looking at weights of the same say 4 students and I am recording it in kilograms, its 68.5 kilograms, 75.3 kilograms, 43.2 kilograms etcetera, I have what is called continuous data. So, we are going to focus on; how do we actually summarize both discrete and continuous data. Again, we will talk about how do we summarize when we have small data that is every discrete data can be treated as a single data value and how do you group data. So, that is the agenda for now.

(Refer Slide Time: 04:57)

Statistics for Data Science -1

- └ Frequency tables
- └ Organizing numerical data

Organizing numerical data

- ▶ Recall, a **discrete variable** usually involves a count of something, whereas a **continuous variable** usually involves a measurement of something.
- ▶ First group the observations into classes (also known as categories or bins) and then treat the classes as the distinct values of qualitative data.
- ▶ Once we group the quantitative data into classes, we can construct frequency and relative-frequency distributions of the data in exactly the same way as we did for categorical data.

So, now let us go ahead and understand how to organize numerical data. Now, a discrete variable involves a count of something, whereas, a continuous variable involves a measurement of something. We have already given example as to what is a discrete data and what is a continuous data. So, the first idea is we already have seen how to summarize the categorical variables using frequency tables.

So, we now want to see whether we can apply the same technique here or the same idea when we want to summarize numerical data. So, one way to do this is to treat the numerical data as categories. So, that can be done by grouping the observations into categories or bins and then treat the classes as distinct values of qualitative data.

Once I have these discrete categories, then I can already know how to construct frequency tables for the categorical data. I can use the same thing and construct a frequency table for my discrete or numerical data. So, let us see how that is done.

(Refer Slide Time: 06:21)

Statistics for Data Science -1
└ Frequency tables
└ Organizing numerical data

Organizing discrete data (single value)

- ▶ If the data set contains only a relatively small number of distinct, or different, values, it is convenient to represent it in a frequency table.
- ▶ Each class represents a distinct value (single value) along with its frequency of occurrence.



So, now suppose my discrete data is a single value data, how do I organize it? So, if the data set consists of relatively small number of distinct different values, then I can treat each value as a category. I repeat if my data set contains only a relatively small number of distinct values, I can treat each distinct value as a category.

And, then once I have each distinct value as a category, I find out what is the frequency of occurrence of each distinct value and I construct a frequency table the way I constructed a frequency table earlier when I talked about categorical data. Let us see this through an example.

(Refer Slide Time: 07:15)



Example

- ▶ Suppose the dataset reports the number of people in a household. The following data is the response from 15 individuals.
 $1, 2, 3, 4, 5$
- ▶ 2,1,3,4,5,2,3,3,4,4,1,2,3,4

Suppose, I have a data set that reports a number of people in a particular household. So, I go to 15 people and I have the following response from 15 individuals. The numbers 2, 1, 3, 4 and 5 and so on indicate the number of people in that particular household. There are 15 values and each one of this value has been picked up from these 15 people. So, now, you can see that I have 5 distinct values of my data and the distinct value of my data are 1, 2, 3, 4 and 5. So, this is the distinct value, I can treat each of these values as a category.

(Refer Slide Time: 08:15)



Example

- ▶ Suppose the dataset reports the number of people in a household. The following data is the response from 15 individuals.
 $2, 1, 3, 4, 5, 2, 3, 3, 4, 4, 1, 2, 3, 4$
- ▶ The distinct values the variable, number of people in each household, takes is 1,2,3,4,5.
- ▶ The frequency distribution table is

Value	Tally mark	Frequency	Relative frequency
1		2	= $2/15$
2		3	= $3/15$
3		5	= $5/15$
4		4	= $4/15$
5		1	= $1/15$
Total		15	1

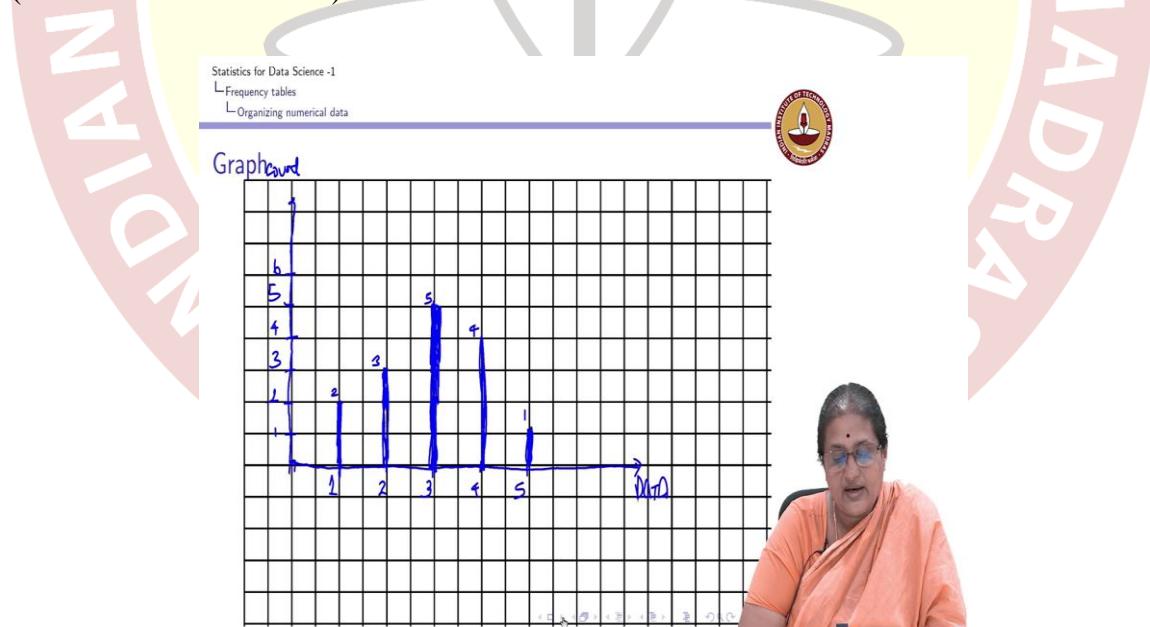


Now, if I treat each one of these values as a category then afterwards I can I put these values. So, you can see in the frequency table I have value 1 here, I have value 2 here, value 3, value 4 and value 5. So, I have these 5 values and for each one of these values I can put a frequency I can construct my frequency distribution table same way I constructed for the categorical variable in my earlier session.

So, now let us construct the frequency table for this data. Now, look at 2. So, I put a tally mark here. 1 is the next value here. So, let us 3, 4, 5, another 2, 3, 3, 3, 4, 4, 1, 2, 3, 4. So, this frequency is 2, this is 3, this is a 5, this is a 4 and this is a 1; this adds up to the 15 which is the number of people we have here. So, we can see that each I can construct the frequency table I have considered each one of these distinct values to be a category.

The relative frequency as we have defined earlier is $2/15$, $3/15$, $5/15$, $4/15$ and $1/15$; we know the relative frequency adds up to 1. So, this is how we can when my data is relatively small in number of distinct values, I can treat each one of the distinct value as a separate category and construct the frequency table just the way we did it for a categorical variable.

(Refer Slide Time: 10:29)



Now, given this I can plot this again, we have seen how we plotted the data. So, I can plot this data in the following way. I have my y-axis, I have my x-axis here. So, let me give the count on the y-axis. The count on the y axis is 1, 2, 3, 4 and 5, 6. I have my data it is 1, 2, 3, 4 and 5 and we can see that, the data frequency the count of value 1 is 2, I

have a bar which is of height 2 here; count of 2 is 3, I have a bar which is of height 3 here; count of 3 is 5, I have a bar which is of height 5 here; count of 4 is 4; count of 5 is 1. So, you can see that, this is very similar to what 2, 3, 5, 4, 1.

So, I can annotate these bars 2, 3, 5, 4 and 1. This is how I can graphically describe the data. What I want you to notice here is, now because this is numerical there is an order in the data. We also described when we talked about plotting bar charts for ordinal data; it is good to preserve the order and data.

So, it is good to have an order here for the data which I have which is number of people in a household. This is the count and this is the data. So, a simple bar chart is what we have shown here. And, since these are distinct values, I am not connecting the bar charts together, I have just listed out what is the height of each bar chart.

(Refer Slide Time: 12:43)



Organize the data into a number of classes to make the data understandable. However, there are few guidelines that need to be followed. They are

1. Number of classes: The appropriate number is a subjective choice, the rule of thumb is to have between 5 and 20 classes.
2. Each observation should belong to some class and no observation should belong to more than one class.

The next thing we want to understand is what would happen when I have continuous data. We said that when I have simple or discrete values which are smaller number then I can treat each discrete value as a category and I can come up with a frequency table. Now, suppose I do not have a small number of discrete values and I have a whole array of values then my; suppose we have a whole array of distinct values then by constructing a frequency table for say I have 65 distinct values, it would be very cluttered.

So, one way to come up with the frequency table for this kind of data is to try and see whether we can group this data into what we refer to as classes. So, now, how do I organize and this can be done even when my data is continuous, ok. So, how do I organize my data into classes?

So, the guidelines we need to follow are the first thing is then I need to choose on how many classes I need. The number of classes is an subjective choice usually, it is good to have between 5 and 20 classes. So, that you do not clutter your table too much. The next thing which we need to understand is each observation and this is very very important; each observation should belong to some class and no observation should belong to more than one class.

What do I mean by the; suppose I have constructed a class 1 and class 2, we will learn now in very shortly we will learn how to construct these class and I have observations 38, 49, 52. I cannot have 38 which belongs to both class 1 and class 2. So, every observation should belong to either class 1 or class 2 and it cannot belong to more than one class, it should belong to one of these classes. So, I could have 38 here, I could have 49 and 52 in the second class. So, every observation is in a class I have defined that is what this means.

(Refer Slide Time: 15:15)

Statistics for Data Science -1
└ Frequency tables
└ Organizing numerical data

Organizing continuous data



Organize the data into a number of classes to make the data understandable. However, there are few guidelines that need to be followed. They are

1. Number of classes: The appropriate number is a subjective choice, the rule of thumb is to have between 5 and 20 classes.
2. Each observation should belong to some class and no observation should belong to more than one class.
3. It is common, although not essential, to choose class intervals of equal length.



And, it is also common though it is not essential, but it is a good practice, at least for beginners to start with class intervals which are of equal length.

(Refer Slide Time: 15:29)

Some new terms



$[30, 40)$
30, 35, 40

1. Lower class limit: The smallest value that could go in a class.
2. Upper class limit: The largest value that could go in a class.
3. Class width: The difference between the lower limit of a class and the lower limit of the next-higher class.
4. Class mark: The average of the two class limits of a class.
5. A class interval contains its left-end but not its right-end boundary point.



So, now let us look at defining a few terms which would be helpful for us to construct frequency tables for group data. So, now, what do I mean? I said that we are going to construct what are called classes or class intervals because of looking at individual distinct values, now I am going to define an interval of values as a category.

So, the minute I define an interval I have what is called a lower class limit. The lower class limit is the smallest value that will get into a class. I have an upper class limit; an upper class limit is the largest value that will go into a class and the class width is the difference between the upper class limit and the lower class limit.

The class mark is the average of the 2 class limits. And, the convention we are going to use is the class interval contains its left end, but not the right end of the boundary point. Now, what do I mean by? Suppose, I have a class interval I have defined as 30 to 40 and my data has 30, 35; 40.

30 and 35 will belong to this class interval, whereas, 40 will not belong to this class interval. I repeat. If I have a class interval 30 to 40, the points 30 and 35 would belong to the class interval, whereas, 40 will not belong to this class interval because it is a right end boundary point of the class interval.

(Refer Slide Time: 17:07)



Example

- The marks obtained by 50 students in a particular course.
- 66, 79, 36, 66, 38, 70, 61, 47, 58, 60, 60, 45, 61, 60, 59, 45, 39, 80, 59, 62, 49, 76, 54, 60, 53, 55, 62, 58, 67, 55, 86, 56, 63, 64, 67, 50, 51, 78, 56, 62, 57, 69, 58, 52, 42, 66, 42, 56, 58.

Class interval	Tally mark	Frequency	Relative frequency
30-40			
40-50			
50-60			
60-70			
70-80			
80-90			
Total			



So, now let us look at an example and see how to construct a frequency table using this example. So, you can see that this has the marks of 50 students in a particular course. You can see that there are lot of distinct values and there are a few values which appear more than once.

But however, the 50 students so I even if I look at the distinct values and try and adopt what I did earlier for a smaller data set my frequency table is going to get very cluttered. So, what I choose is I am going to choose class intervals of size 10 and start putting in each value after choosing the class interval of size 10.

So, the class interval size once I having defined my class intervals, I see that the minimum occurs in the range of 30. So, I start with a minimum of 30 and go up to 90, I do not have a mark beyond 90 and these are my class intervals. Now, I am assuming each class interval is of the same size. Hence, my class intervals are 30 to 40, 40 to 50 and so on 80 to 90. So, the assumption I have here is each of my class intervals are of the same size.

So, I, now these are my categories and I am going to put in each of the values from here into one of these intervals. Again remember, every value here goes into one of these buckets and no value will go into two of these buckets. I include for example, in this 30 to 40 I include 30, but exclude 40.

So, 68 goes here, 79 goes here, 38 goes here, 68 again goes here, 35 goes here, now come to 70; 70 goes here and because 70 is excluded from this interval, 70 goes here, 61 here, 47 here, 58 goes here, 66 goes here, 60 again comes here, 45 goes here, 61 comes here, 60 again comes here, 59 is here, 45 is here.

(Refer Slide Time: 19:43)

Statistics for Data Science -1
 └ Frequency tables
 └ Organizing numerical data

Frequency table

68, <u>79</u> , <u>38</u> , 68, <u>35</u> , <u>70</u> , 61, 47, 58, 66, 60, 45, 61, 60, 59, 45, <u>59</u> , <u>80</u> , 59, 62, 49, <u>76</u> , 54, 60, 53, 55, 62, 58, 67, 55, <u>86</u> , 56, 63, 64, 67, 50, 51, <u>78</u> , 56, 62, 57, 69, 58, 52, 42, 66, 42, 56, 58.			
Class interval	Tally mark	Frequency	Relative frequency
30-40		3 ✓	0.06
40-50		6	0.12
50-60		18	0.36
60-70		17	0.34
70-80		4 ✓	0.08
80-90		2 ✓	0.04
Total		50	1

So, if I continue this I get a frequency table which is of this kind. So, you can see that, 30 to 40 you can easily see its 38, 35 and 39 these are the 3 things. So, it has a frequency of 3, 80 to 90 has 86, 80; because 80 to 90 is in that it has a frequency of 2. 70 to 80 has 79, 70, 76, 78 ok, with the frequency 4 and all these frequencies as always add up to 50, the relative frequencies are also given here.

Now, remember here what we have done and what is the difference between what we have done here and the earlier part is we have grouped the data into class intervals. So, this is how we construct a frequency table when I have group data and this could be even for continuous data I can use the same way.

(Refer Slide Time: 20:57)



1. Frequency table for discrete single value data.
2. Frequency table for continuous data using class intervals.



So, now what we have learned so far is how to construct a frequency table for a single value discrete single value data and how to construct a frequency table for continuous data using class intervals.

(Refer Slide Time: 21:17)



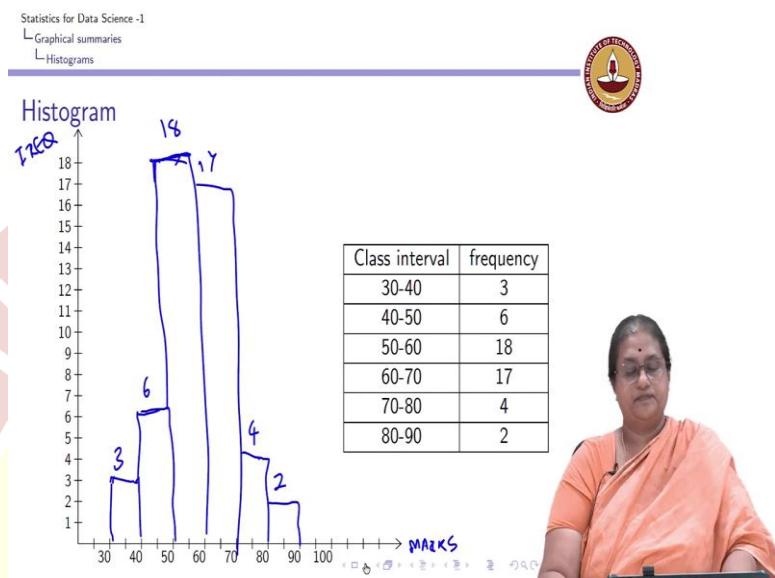
- Step 1 Obtain a frequency (relative-frequency) distribution of the data.
- Step 2 Draw a horizontal axis on which to place the classes and a vertical axis on which to display the frequencies (relative frequencies).
- Step 3 For each class, construct a vertical bar whose height equals the frequency (relative frequency) of that class.
- Step 4 Label the bars with the classes, the horizontal axis with the name of the variable, and the vertical axis with "Frequency" ("Relative frequency").



Now, we will go about seeing now how do I summarize the data for the second case that is where I have constructed a frequency table for continuous data. The way we do it is through what we call histogram. Histogram is one of the most popular graphical summary of a continuous data. So, now, we will understand how to set up a histogram.

Now, when we want to set up a histogram the first thing is we obtain a frequency distribution of data. We have already seen how to obtain this frequency distribution of data. Draw horizontal axis.

(Refer Slide Time: 22:01)



So, on the horizontal axis I have laid my 30 to 40, 40 to 50 and these are my marks that is on my horizontal axis. A vertical axis on which you display frequency. So, on this axis I have the frequency, on this axis I have my marks which are given in the as intervals. The next thing is for each class construct a vertical bar whose height equals the frequency. So, now, for each class, so you can see for 30 to 40 I have a frequency 3; so, I construct a bar with frequency 3.

40 to 50 I have frequency 6; so, I construct a bar with frequency 6. 50 to 60, its 18, I construct a bar with frequency 18 then I come its frequency 17. 70 to 80 is frequency 4 and the last thing is frequency 2. So, this you can annotate it. I have 3, I have 6, I have 18, these are my counts, this is 17, this is 4 and this is 2. So, I have constructed my histogram.

Notice the difference between a histogram and the bar chart, because my class intervals are in a sense continuous. I am not leaving a gap between these bars. So, it is a continuous display of data. The vertical height of this bar represents the count in every class interval. So, this is how we have. So, I label the bars, the vertical axis and this is how we construct a histogram.

(Refer Slide Time: 24:11)

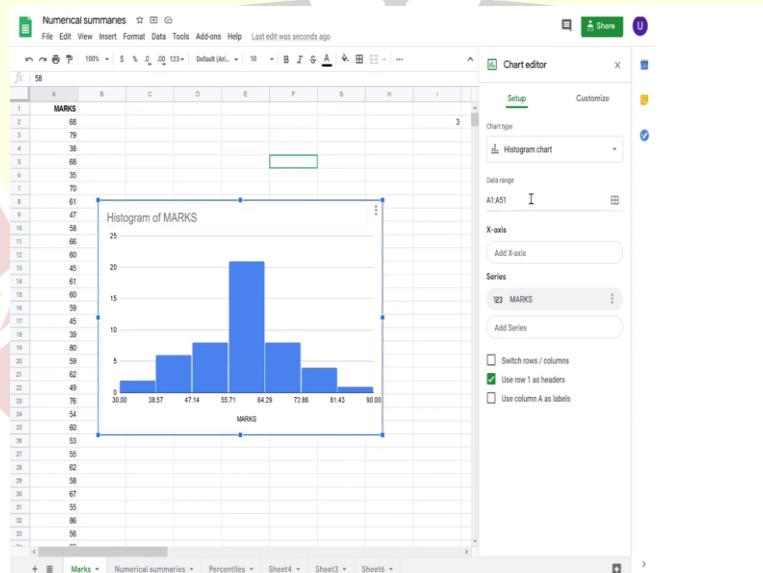
The screenshot shows a Google Sheets interface with a histogram titled "Histogram". The chart displays the frequency distribution of marks for 50 students. The x-axis represents marks from 30.00 to 90.00, and the y-axis represents frequency from 0 to 25. The data is divided into bins: [30.00, 38.57], [38.57, 47.14], [47.14, 55.71], [55.71, 64.29], [64.29, 72.86], [72.86, 81.43], and [81.43, 90.00]. The frequencies are approximately 2, 5, 8, 20, 10, 4, and 1 respectively.

[https://docs.google.com/spreadsheets/d/109W3ga8TZG3pWJwofG4h0yE7xvoGOK_kCvmm0e9w0kQ/edit?
usp=sharing](https://docs.google.com/spreadsheets/d/109W3ga8TZG3pWJwofG4h0yE7xvoGOK_kCvmm0e9w0kQ/edit?usp=sharing)



So, I can construct a histogram using my Google Sheets as well and that is what we are going to discuss now.

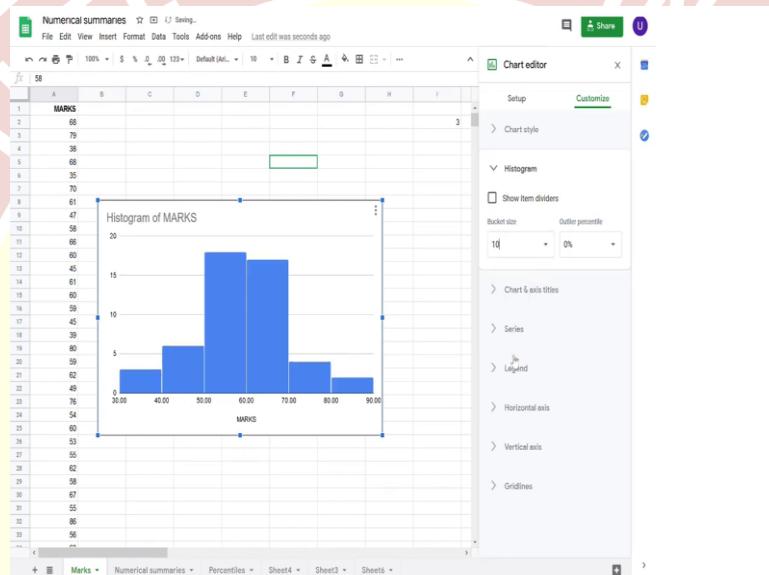
(Refer Slide Time: 24:19)



So, this is the data which we have. So, the same data which was for 50 people is listed here. So, I have a data for 50 students, which is listed here. You can see the data which is here. So, the data is for 50 students. This is the same data which we had here. This is given in column A. The first row is the heading for the row.

So, how do I construct histogram using my Google Sheet? Go to Insert, in Insert you have a Chart. Now, in this chart you can choose the data. So, you can see that the chart type is a Histogram chart. The data range is your A2 to A1 to A51, but I have clicked on using row 1 as headers that is being clicked here. Using row 1 as a header is being clicked, ok. And, the series is the marks. So, you can see that the minute I do not specify the equal buckets here, the histogram randomly chooses the end points of each class interval.

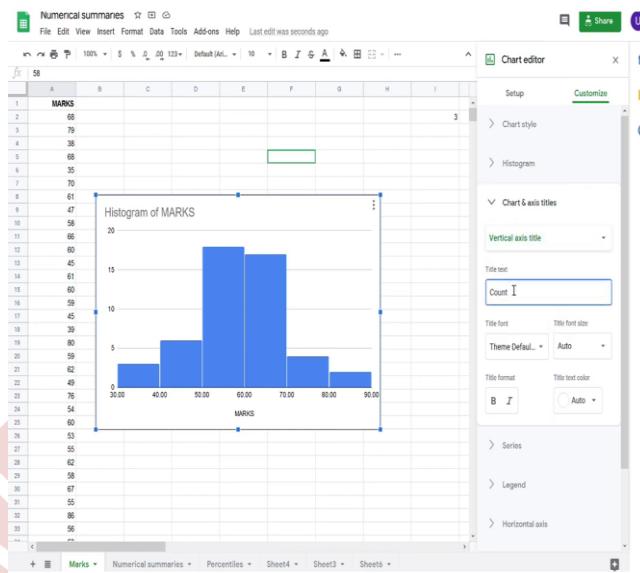
(Refer Slide Time: 25:55)



So, I go to customize, in the customize thing, I go to chart in under my histogram, there is a heading which is called bucket size. In bucket size, I click on 10, that is the size of my class interval instead of auto and you can see that this is precisely what we earlier drew freestyle with 30 to 40 having a frequency of 3 and this is 18, this is 17, this is 4 and this is 2.

As again as always we can write down what are the legends, we can find out what is the what is my horizontal axis, what is your; you can go to the chart, you can have what is your title text, you can find out what is your subtitle, what is your horizontal access title, I can list it as marks.

(Refer Slide Time: 26:43)



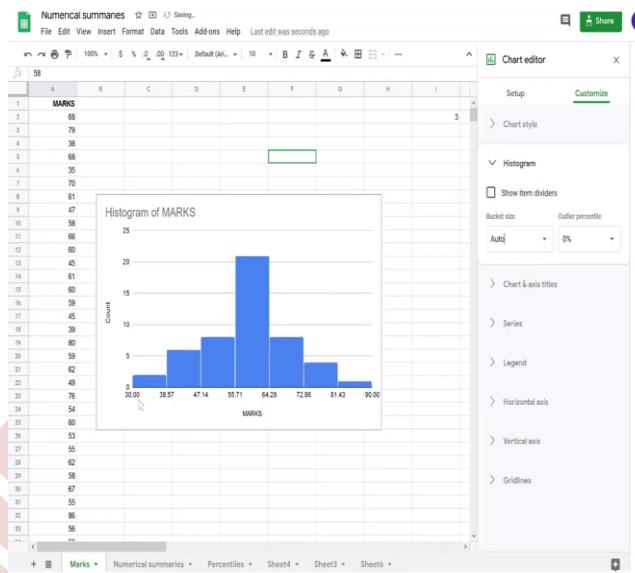
My vertical axis title I can list it as count and that is what I have the count and I can remove grid lines and I can do everything else, but this is how I construct a histogram using my Google Sheets.

(Refer Slide Time: 27:01)

The screenshot shows a presentation slide titled "Histogram" from a course on "Statistics for Data Science - 1". The slide displays a histogram titled "Distribution of marks" with the vertical axis labeled "Frequency" ranging from 0 to 20 and the horizontal axis labeled "Marks" ranging from 30.00 to 90.00. The histogram bars are blue. Below the histogram, there is a URL: https://docs.google.com/spreadsheets/d/109W3ga8TZG3pWJwofG4h0yE7xvoGOK_kCvmm0e9w0kQ/edit?usp=sharing. The slide also features the Indian Institute of Technology Madras logo.

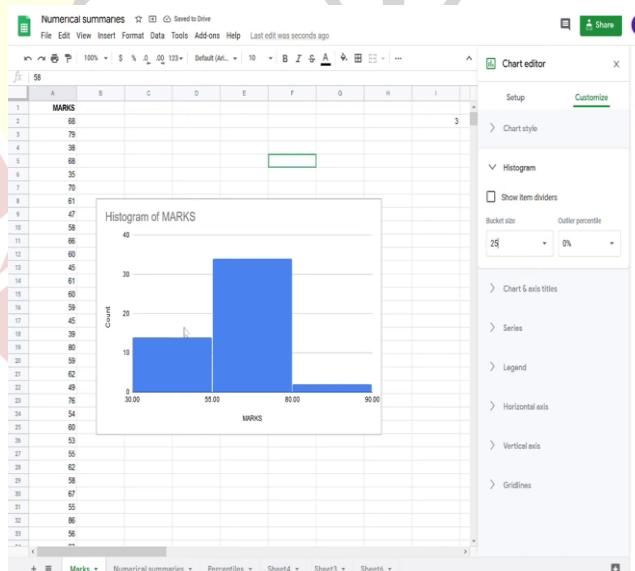
So, this is what we have. And, this is how we showed. We construct a histogram using the Google Sheets.

(Refer Slide Time: 27:17)



Again, go back and see that if I what I specified here is the bucket size of 10. If you list auto, it will create between 30 and 90 it decides on what is the appropriate class interval. But here I want a bucket size of 10.

(Refer Slide Time: 27:33)



So, you can see that if you increase the bucket size, it is not giving a proper. So, the size of your class interval it actually matters and it is good to have a reasonable size and in this problem I have chosen a bucket size of 10.

(Refer Slide Time: 27:53)

Statistics for Data Science -1
└ Graphical summaries
└ Stem-and-leaf diagram

Stem-and-leaf diagram

1/5



Definition

In a stem-and-leaf diagram (or stemplot)¹, each observation is separated into two parts, namely, a stem-consisting of all but the rightmost digit-and a leaf, the rightmost digit.

- For example, if the data are all two-digit numbers, then we could let the stem of a data value be the tens digit and the leaf be the ones digit.

¹Weiss, Neil A. Introductory Statistics: Pearson New International Edition.
Pearson Education Limited, 2014.



The next important graphical summary which we are going to discuss is, popularly referred to as a stem and leaf diagram. In a stem and leaf diagram also popularly referred to as a stem plot each observation is separated into 2 parts; namely a stem consisting of all but the right most digit and a leaf which has the right most digit.

Now, what do we mean by a right most digit and a left most digit? So, in a stem plot for example, if I have 2 digit numbers then we could let the stem of the data to be the value of the tens and the leaf to be the ones. For example, if I have 15 the stem would be 1 and the leaf would be 5 that is what I mean by a stem of a data and leaf of a data.

(Refer Slide Time: 28:59)

Statistics for Data Science -1
└ Graphical summaries
└ Stem-and-leaf diagram

Stem-and-leaf diagram



Definition

In a stem-and-leaf diagram (or stemplot)¹, each observation is separated into two parts, namely, a stem-consisting of all but the rightmost digit-and a leaf, the rightmost digit.

- For example, if the data are all two-digit numbers, then we could let the stem of a data value be the tens digit and the leaf be the ones digit.

- The value 75 is expressed as

Stem	Leaf
7	
5	

- The two values 75, 78 is expressed as

Stem	Leaf
7	
5, 8	

¹Weiss, Neil A. Introductory Statistics: Pearson New International Edition. Pearson Education Limited, 2014.



The value 75 can be represented as 7 and 5. Now, if I have 2 values then for 75 stem was 7, leaf was 5. For 78 the stem is again 7 and leaf is 8 and I can represent this value by just having 1 stem and 2 leaves. So, this is called a stem and leaf plot.

(Refer Slide Time: 29:33)

Statistics for Data Science -1
└ Graphical summaries
└ Stem-and-leaf diagram

Steps to construct a stemplot



Step 1 Think of each observation as a stem—consisting of all but the rightmost digit—and a leaf, the rightmost digit.

Step 2 Write the stems from smallest to largest in a vertical column to the left of a vertical rule.

Step 3 Write each leaf to the right of the vertical rule in the row that contains the appropriate stem.

Step 4 Arrange the leaves in each row in ascending order.



A stem and leaf plot how do I construct it? So, look at any observation as a stem consisting of all but the right most digit and a leaf has the right most digit. Write the stems from the smallest to the largest in a vertical column to the left of the rule. Write

each leaf to the vertical and arrange the leaves. This is how we construct a stem plot we apply that to an example.

(Refer Slide Time: 29:59)



Example

- The following are the ages, to the nearest year, of 11 patients admitted in a certain hospital: 15, 22, 29, 36, 31, 23, 45, 10, 25, 28, 48 ✓

STEM	LEAF
1	0, 5
2	2, 3, 5, 8, 9
3	1, 6
4	5, 8



I have the ages to the nearest year of 11 patients admitted in a certain hospital. The ages are 15, 22, 29, 36, 31, 23, 45, 10, 25, 28 and 14. So, let us go back and see; what are these steps. The first is think each observation as a stem and a leaf. Write the stems from smallest to largest. So, if I write the stem the smaller stem is 1, 2, 3 and 4.

These are my stems ok; these are my stems. Now, this is my first step. I have second step where I have written the stems from the smallest to largest in a vertical column. Write each leaf to the right of the vertical rule. So, if I come here these are my leaves. 5 comes here, 2 comes here. So, let me 9 is again here, 6 is here, 1 is here, 23, 3 is here, 45 is here, 10 0 is here, 25 is here, 28 is here and 48 is here.

So, this is what I have done in step 3. We have written each leaf to the right of the vertical rule. The last step is arrange the leaves in ascending order. So, the way I can arrange the leaf here is I get a 0 and 5 that is the arrangement. Here this is going to be 2, 3, 5; 8 and 9; this would be 1 and 6 and the last is going to remain as it is to have a final stem plot as the following.

(Refer Slide Time: 32:15)

Statistics for Data Science -1
└ Graphical summaries
 └ Stem-and-leaf diagram

Example



- ▶ The following are the ages, to the nearest year, of 11 patients admitted in a certain hospital: 16, 22, 29, 36, 31, 28, 45, 10, 25, 26, 48
- ▶ Draw a stem-and-leaf plot for this data set.

1	05
2	23589
3	16
4	58



Which is 1; 0, 5; 2; 2, 3, 5, 8, 9, 3; 1, 6, and 4; 5, 8. 10 corresponds to 10, 15 corresponds to 15 then I have a 22, I have a 23, I have a 25, I have a 28, a 29, 31, 36, 45 and 48. So, this is how we construct a stem plot and we can get a stem plot many of the statistical packages give you a stem plot.

(Refer Slide Time: 32:57)

Statistics for Data Science -1
└ Graphical summaries
 └ Stem-and-leaf diagram

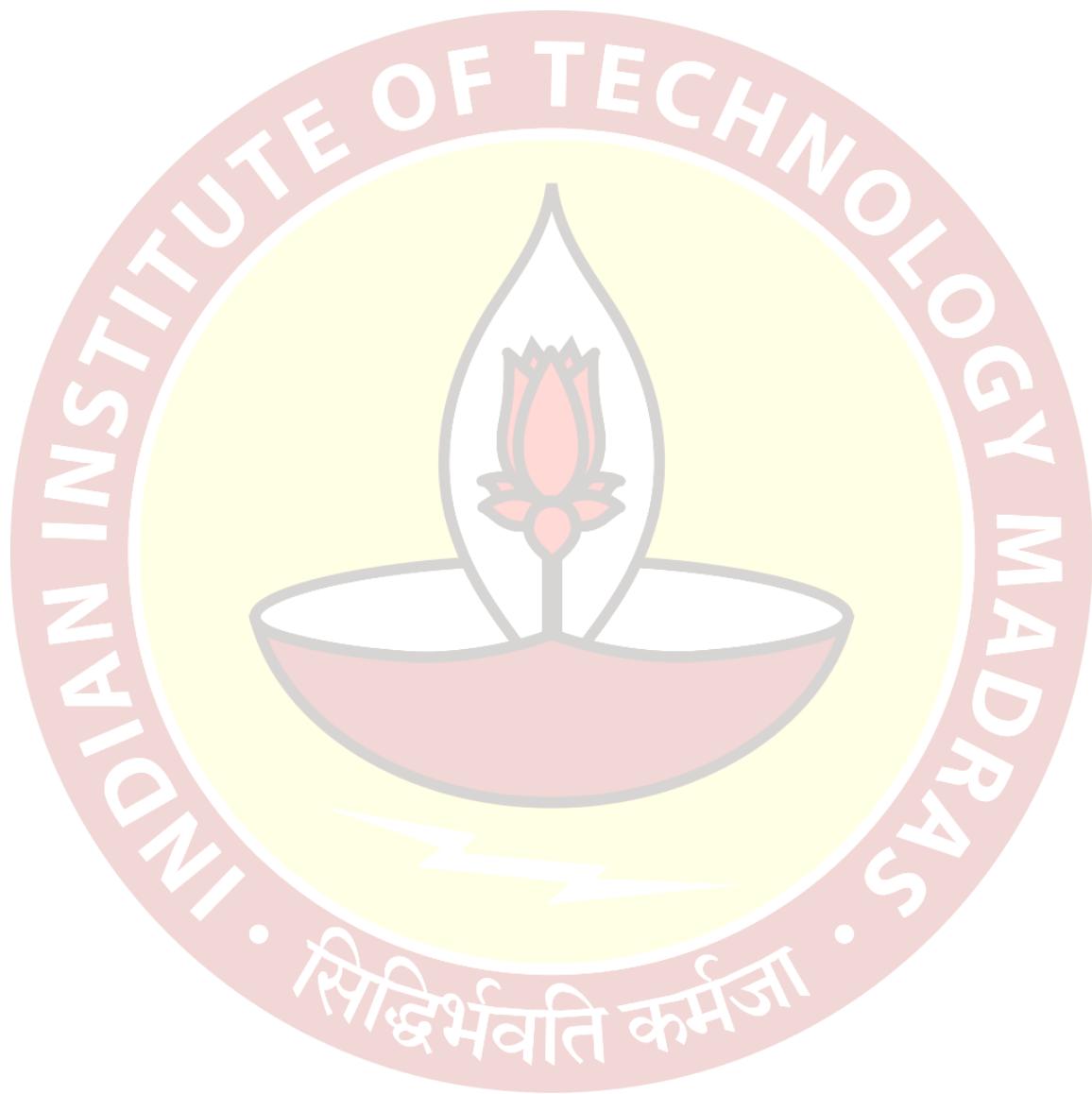
Section summary



1. Construct a histogram for grouped data.
2. Construct a stemplot to describe numerical data.

So, what we have learnt in this section so far is, how do we construct a histogram. First of all we learned as to how do we construct a frequency table both when I have single value discrete values are very small number of data. And, then when I have large number

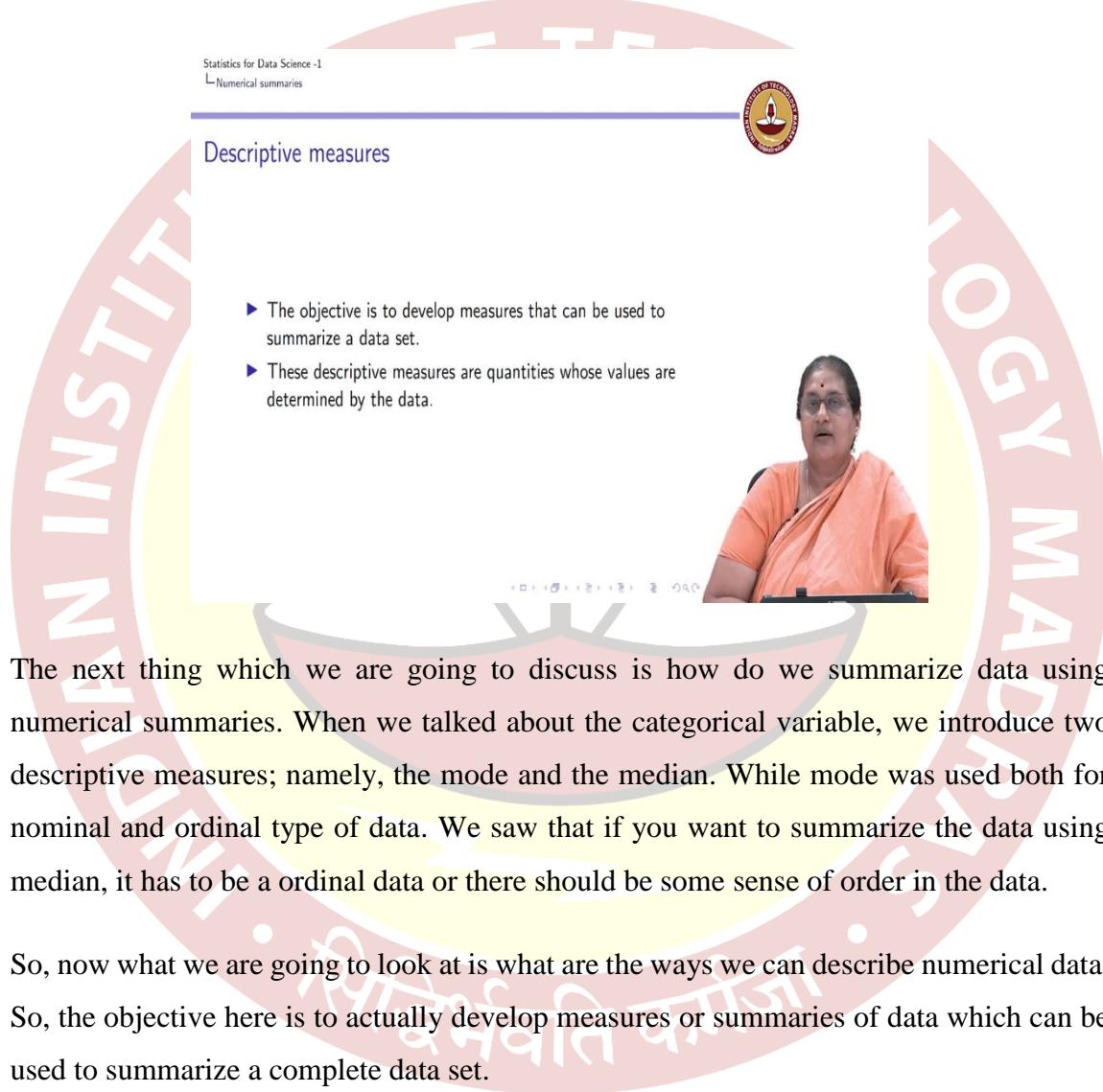
of values where I group the data. And, once I have the frequency tables for my group data, we saw first how to construct a histogram by constructing class intervals of equal length and constructing a stem leaf plot for my data. This is what we have learnt so far.



Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 3.2
Describing Numerical Data – Mean

(Refer Slide Time: 00:14)



Statistics for Data Science -1
↳ Numerical summaries

Descriptive measures

The objective is to develop measures that can be used to summarize a data set.

These descriptive measures are quantities whose values are determined by the data.

The next thing which we are going to discuss is how do we summarize data using numerical summaries. When we talked about the categorical variable, we introduce two descriptive measures; namely, the mode and the median. While mode was used both for nominal and ordinal type of data. We saw that if you want to summarize the data using median, it has to be a ordinal data or there should be some sense of order in the data.

So, now what we are going to look at is what are the ways we can describe numerical data. So, the objective here is to actually develop measures or summaries of data which can be used to summarize a complete data set.

(Refer Slide Time: 01:07)

Descriptive measures



Most commonly used descriptive measures can be categorized as

- ▶ **Measures of central tendency:** These are measures that indicate the most typical value or center of a data set.
- ▶ **Measures of dispersion:** These measures indicate the variability or spread of a dataset.



So, the most commonly used, so especially when we are dealing with numerical data the most commonly used descriptive measures again can be broadly categorized into two categories and these categories are one which are known as measures of central tendency and the second which are known as measures of dispersion.

Now, what do we mean by measures of central tendency? Measures of central tendency as the name suggests, it talks about where the data is actually concentrated or what is the most typical value of a data set. A measure of central tendency describes or tells us what we can expect as a typical value of a data set; whereas, the measure of dispersion also referred to as measures of variation or measures of spread talk about the variability in a data set or spread in a data set.

(Refer Slide Time: 02:15)

The most commonly used measure of central tendency is the mean.

Definition

The *mean* of a data set is the sum of the observations divided by the number of observations.

- ▶ The mean is usually referred to as *average*.

- ▶ Arithmetic average; divide the sum of the values by the number of values (another typical value)

- ▶ For discrete observations:

$$\text{Sample mean: } \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

n
SAMPLE SIZE
 N
Population size



So, we will go through these measures in detail. When we talk about measures of central tendency, the most common measure of central tendency is what we refer to as the mean of a data set.

What is a mean? I define the mean of a data set to be the sum of the observations divided by the number of observations. So, now, let us formally define what it is. Suppose, I have x_1, x_2, x_3 , I refer to my observations as x_1, x_2, x_3 . For example, I have my data which is 4, 3, 1, 2 and 5. I have n equal to 5 here. My x_1 is 4; x_2 is 3; x_3 is 1; x_4 is 2 and x_5 is 0.

So, in general, if I have n observations each x_1 refers to the first observation in my data set; x_2 refers to the second observation in my data set and x_n refers to the n th observation in my data set. The mean of a data set is the sum of these observations, the numerator gives me the sum of these observations which is nothing but $x_1 + x_2 + x_3 \dots + x_n$ divided by the total number of observations which is n , small n .

Now, recall in when we introduce the notion of a sample and population, we said that a small n refers to what we call the sample size; whereas, capital N refers to a population size. So, if I am having a data set which is a sample, then I do denote it or my notation is going to be small n ; when I have population, when I refer to it as a population, it is going to be capital N .

So, if my I can define a sample mean, when I have a sample data set to be the sum of the sample observations divided by the total number of observations, the way we have done here.

(Refer Slide Time: 04:46)



The Mean

The most commonly used measure of central tendency is the mean.

Definition

The mean of a data set is the sum of the observations divided by the number of observations.

- ▶ The mean is usually referred to as average.
- ▶ Arithmetic average; divide the sum of the values by the number of values (another typical value)
- ▶ For discrete observations:
 - ▶ Sample mean: $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$
 - ▶ Population mean: $\mu = \frac{x_1 + x_2 + \dots + x_n}{N}$



I can also define a population mean which is typically refer to by the Greek alphabet μ to be $x_1 + x_2 + \dots + x_n$; whereas, again remember, I said n is the total number of observations in a population divided by capital N which is the total number of observations.

The definitions are the same, only thing the number of observations differ whether you are referring to a sample or whether you are referring to a population. The mean is also popularly refer to as a average. So, now, let us compute the mean for small data sets.

(Refer Slide Time: 05:26)



$$1. \text{Data set: } 2, 12, 5, 7, 6, 7, 3; \bar{x} = \frac{2+12+5+7+6+7+3}{7} = \frac{42}{7} = 6$$
$$2. \text{Data set: } 2, 105, 5, 7, 6, 7, 3; \bar{x} = \frac{2+105+5+7+6+7+3}{7} = \frac{135}{7} = 19.285$$



Now, I have this data set which is given to me. So, the mean for this data set is very simple. That is \bar{x} is $2 + 12 + 5 + 7 + 6 + 7 + 3$. This is what I have here and that I can see is equal to 42, the numerator and which is equal to 6. In other words, 6 is the mean of this data set. Now, let us look at another data set.

See the only difference between the first data set and the second data set is in x_2 or the second observation, everything else is the same. So, I compute the mean for this data set, When I compute the mean for this data set, I find \bar{x} is equal to 135 by 7 which is 19.285. What I want you all to observe very clearly is the difference between this data set and this data set is only the second number; but the mean difference is very high.

(Refer Slide Time: 06:52)

Example



1. 2, 12, 5, 7, 6, 7, 3;
 $\bar{x} = \frac{2+12+5+7+6+7+3}{7} = \frac{42}{7} = 6$
2. 105, 5, 7, 6, 7, 3 $\bar{x} = \frac{2+105+5+7+6+7+3}{7} = \frac{135}{7} = 19.285$
3. 2, 105, 5, 7, 6, 3 $\bar{x} = \frac{2+105+5+7+6+3}{6} = \frac{128}{6} = 21.33$



Now, let us look at another data set which is the following, where it has only 6 observations now, it does not have the last observation. The mean for this data set is going to be 21.33. We see that these two means are fairly close to each other; whereas, these two means are very different from each other.

So, what this tells us is the mean even though here I have only one observation which is different. This observation and this observation are very different from each other. So, the mean is very sensitive to outliers. By outliers, I mean what is the number which is very different from what the typical data set behaves like.

(Refer Slide Time: 07:48)

Statistics for Data Science - I
└ Numerical summaries
 └ Measures of central tendency

Example

► The marks obtained by ten students in an exam is
68, 79, 38, 68, 35, 70, 61, 47, 58, 66

► The sample mean is

$$\frac{68 + 79 + 38 + 68 + 35 + 70 + 61 + 47 + 58 + 66}{10} = \frac{590}{10} = 59$$


↓

- The marks obtained by ten students in an exam is
68, 79, 38, 68, 35, 70, 61, 47, 58, 66
- The sample mean is

$$\frac{68 + 79 + 38 + 68 + 35 + 70 + 61 + 47 + 58 + 66}{10} = \frac{590}{10} = 59$$



File Edit View Insert Format Data Tools Help Last edit was 16 minutes ago

25

So, now let us go back to another example. These are the marks obtained by 10 students in an exam. So, now, I can see that the sample mean in this case is 590 divided by 10 which is 59.

(Refer Slide Time: 08:07)

	A	B	C	D
1		DATA	Adding constant	Multiplying with cons
2		68	73	
3		79	84	
4		38	43	
5		68	73	
6		35	40	
7		70	75	
8		61	66	
9		47	52	
10		58	63	
11		66	71	
12	TOTAL	590	640	
13	MEAN	59	64	
14	MEDIAN	63.5	68.5	
15	MODE	68	73	
16	VARIANCE	210.8888889	210.8888889	33.74222

So, this is the data I have. So, these are the 10 data sets. This is the data which I have, which I have just given to you 68, 79; this is the data which we have worked with. So, in this data, you can see that this sum $B2$ to $B11$ is nothing but the sum which I have worked here.

So, the sum of this is the numerator sum which is equal to 590. I have a 590 here which is the sum of all my data values and you can see the mean, the google sheet the mean gives me what is $B12$ by 10 that is what is 590 divided by the total number of observations.

(Refer Slide Time: 09:06)

A screenshot of a Google Sheets document titled 'Numerical summaries'. The data is as follows:

	A	B	C	D
9		47	52	
10		58	63	
11		66	71	
12	TOTAL	590	640	
13	MEAN	59	64	
14	MEDIAN	63.5	68.5	
15	MODE	68	73	
16	VARIANCE	210.8888889	210.8888889	33.74222
17	N	10	10	
18	n	10	10	
19	Population variance	189.8	189.8	30.
20	Sample variance	210.8888889	210.8888889	33.74222
21	Standard deviation	14.52201394	14.52201394	5.808805
22		14.52201394	14.52201394	5.808805
23		59		
24				

That is what the mean gives you. Now, you can see that the google sheet has average demand, command is giving me the same average as what I have obtained before. So, average command in the google sheet gives me the mean which we have different ok. The next thing which we are going to see is how do I obtain the mean for a grouped data.

(Refer Slide Time: 09:36)

Statistics for Data Science -1

- Numerical summaries
- Measures of central tendency

Mean for grouped data: discrete single value data

► The following data is the response from 15 individuals.
 $\begin{array}{c} 2, 1, 3, 4, 5, 2, 3, 3, 3, 4, 4, 1, 2, 3, 4 \end{array}$

►
$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \dots + f_nx_n}{n}$$

Value(x_i)	Tally mark	Frequency(f_i)	$f_i x_i$
1		2	2
2		3	6
3		5	15
4		4	16
5		1	5
Total		15	44

Mean = $\frac{\sum f_i x_i}{n} = \bar{x}$

Mean =

Now, if I have single. So, now, let us go back to this data which we have seen earlier which was the response from 15 individuals, we have already seen that this is the frequency 1 appears twice; 2 appears thrice; 3 appears five times; 4 appears four times and 5 appears once, this is what we have already seen.

Now, for a data of this kind how do I compute the average? Now, 1 appears two times, so it appears with the frequency 2. So, the way I need to add 1 2 times. So, it is 2×1 so $1 + 1$ which is giving me a 2. Similarly, 2 appears 3 times. So, the total sum of 2 is going to be $2 + 2 + 2$ or 2×3 which is equal to 6. 3 appears five times, so the sum 3 contributes to its 3×5 which is 15; 4 into 4 which is 16 and 5 into 1 is 15.

So, instead of writing $2 + 1 + 3 +$ all of it, I can write it as 1 appears two times which contributes to 2; 2 appears three times which contributes six and this is the numerator which is $\sum_{i=1}^n f_i x_i$ divided by your n or $\sum f_i$ which is equal to n . This gives me the \bar{x} when I have discrete single value data.

(Refer Slide Time: 11:31)

Mean for grouped data: discrete single value data

INDIAN INSTITUTE OF TECHNOLOGY
MADRAS
 Statistics for Data Science - I
 └── Numerical summaries
 └── Measures of central tendency

- The following data is the response from 15 individuals.
 $2, 1, 3, 4, 5, 2, 3, 3, 3, 4, 4, 1, 2, 3, 4$
- $\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \dots + f_n x_n}{n}$

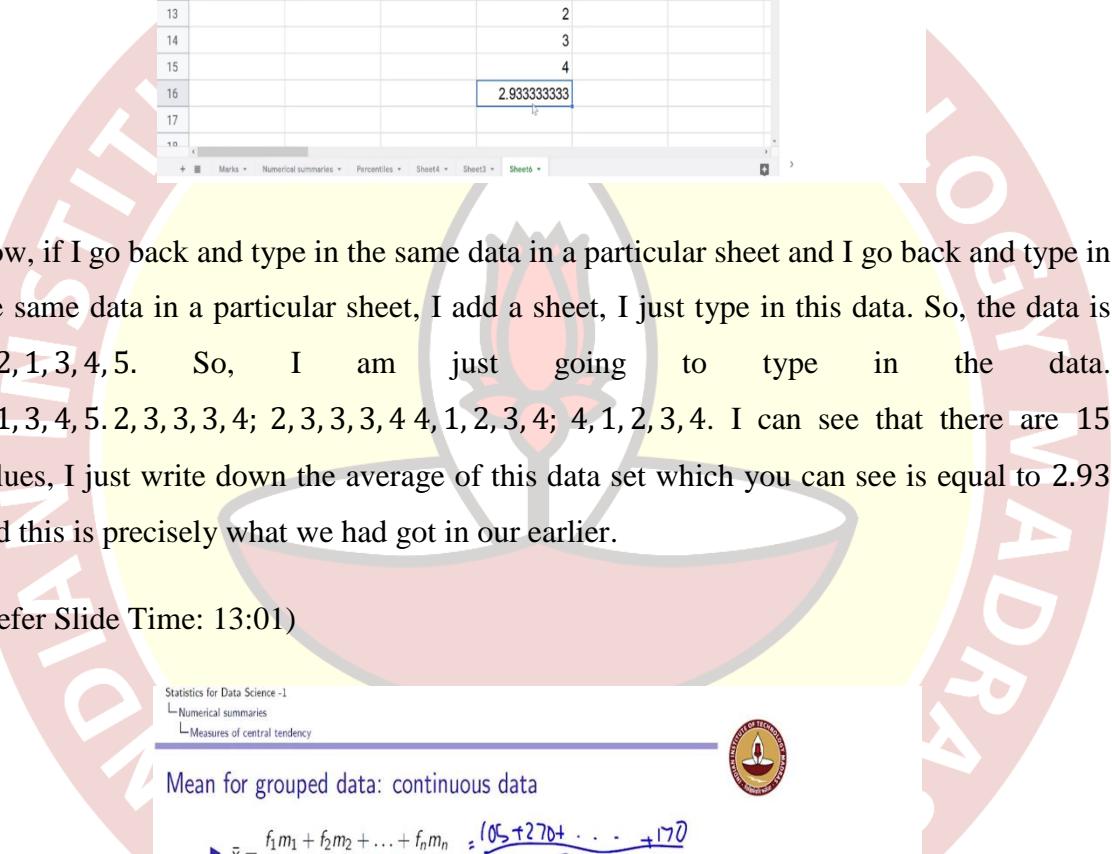
Value(x_i)	Tally mark	Frequency(f_i)	$f_i x_i$
1		2	2
2		3	6
3		5	15
4		4	16
5		1	5
Total		15	44

$\text{Mean} = \frac{44}{15} = 2.93$

26

So, the mean here is going to be 44 divided by 15 which is equal to 2.93.

(Refer Slide Time: 11:55)

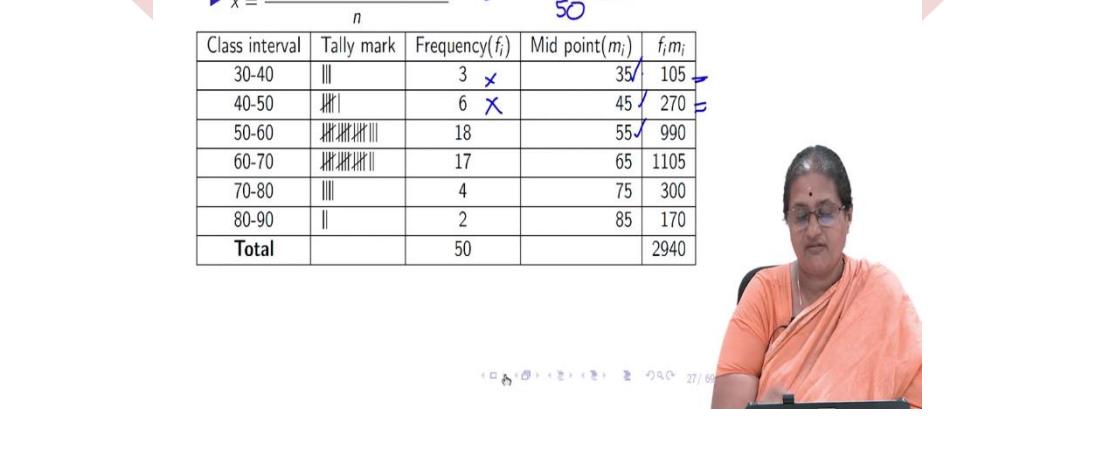


A screenshot of a spreadsheet application showing a table of data. The table has columns A through F. Rows 2 through 10 contain numerical values: 38, 47, 58, 61, 66, 68, 68, 70, and 79 respectively. Row 16 contains the formula `=average(B2:B10)` and the result `2.933333333`. The background features a large circular watermark with the text "DATA SCIENCE" and "LOGIC MASTERS" and a central logo.

A	B	C	D	E	F
2	38			1	
3	47			3	
4	58			4	
5	61			5	
6	66			2	
7	68	68		3	
8	68			3	
9	70			3	
10	79			4	
11				4	
12				1	
13				2	
14				3	
15				4	
16				2.933333333	
17					

Now, if I go back and type in the same data in a particular sheet and I go back and type in the same data in a particular sheet, I add a sheet, I just type in this data. So, the data is 1, 2, 1, 3, 4, 5. So, I am just going to type in the data. 2, 1, 3, 4, 5. 2, 3, 3, 3, 4; 2, 3, 3, 3, 4 4, 1, 2, 3, 4; 4, 1, 2, 3, 4. I can see that there are 15 values, I just write down the average of this data set which you can see is equal to 2.93 and this is precisely what we had got in our earlier.

(Refer Slide Time: 13:01)



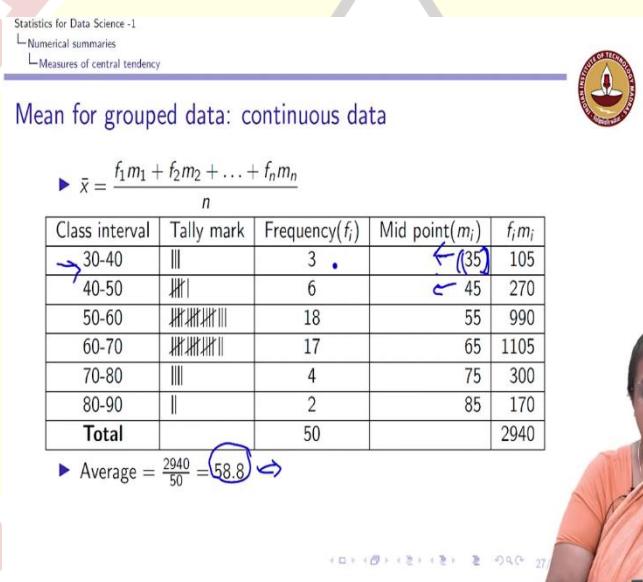
A screenshot of a presentation slide titled "Mean for grouped data: continuous data". It shows the formula for calculating the mean of grouped data: $\bar{x} = \frac{f_1m_1 + f_2m_2 + \dots + f_nm_n}{n}$, with a handwritten note above it showing the calculation $\frac{105+270+\dots+170}{50}$. Below the formula is a table of grouped data:

Class interval	Tally mark	Frequency(f_i)	Mid point(m_i)	$f_i m_i$
30-40		3	35	105
40-50		6	45	270
50-60		18	55	990
60-70		17	65	1105
70-80		4	75	300
80-90		2	85	170
Total		50		2940

Now, what how do we compute the mean for group data or continuous data? Now, again recall this when I have continuous data, I do not have a particular value of the discrete data that is being taken. So, in this case recall we have what is called the midpoint of the data set or for each class interval. For example, the midpoint of the class interval 30 to 40 is 35; 40 to 50 is 45; 50 to 60 is 55. So, you look at the midpoint rather than looking at each discrete value, I know the frequency of the data set.

Now, I multiply this frequency with the midpoint, I get 105 and 6 into 45 is 270. My numerator is going to be 105 plus 270 plus 170 divided by the total number is again 50. So, it is 2940 divided by 50 which is 58.8.

(Refer Slide Time: 14:11)



A word of caution here is this 58.8 is not the actual mean because we are approximating it only with the midpoint, I am now taking the actual values of the data. I repeat this 58.8 is an approximation because I am multiplying the frequency with the midpoint or the best representative in this class interval which is 35. I know my data lies between 30 to 40. So, I am just saying that ok, it is around 35. So, this 35 is an approximation.

(Refer Slide Time: 15:04)

Statistics for Data Science -1

- └ Numerical summaries
- └ Measures of central tendency

Mean for grouped data: continuous data

► $\bar{x} = \frac{f_1m_1 + f_2m_2 + \dots + f_nm_n}{n}$

Class interval	Tally mark	Frequency(f_i)	Mid point(m_i)	$f_i m_i$
30-40		3	35	105
40-50		6	45	270
50-60		18	55	990
60-70		17	65	1105
70-80		4	75	300
80-90		2	85	170
Total		50		2940

► Average = $\frac{2940}{50} = 58.8$.

► 58.8 is an approximate and not exact value of the mean

So, because we are not looking at the exact data values, this average is only an approximation and not the exact value of the mean; whereas, when we looked at the discrete single value data, I took the exact value of my data and hence, my mean matched with the exact average.

(Refer Slide Time: 15:24)

Statistics for Data Science - I

- ↳ Numerical summaries
- ↳ Measures of central tendency

$\bar{x} = \frac{\sum x_i}{10}$ $\bar{y} = \frac{\sum y_i}{10}$

Adding a constant

x_1	x_2	x_3	x_4
68	78	38	66
y_1	y_2	y_3	y_4
73	84	43	71

- ▶ Let $y_i = x_i + c$ where c is a constant then $\bar{y} = \bar{x} + c$
- ▶ Example: Recall the marks of students
68, 79, 38, 68, 35, 70, 61, 47, 58, 66.
 - ▶ Suppose the teacher has decided to add 5 marks to each student.
 - ▶ Then the data becomes
73, 84, 43, 73, 40, 75, 66, 52, 63, 71.

This is a point to be noted. Now, how sensitive is the mean if I add a constant to every observation. Now, why are we even interested in knowing about this? For example, let us look at a case where a teacher has already given the marks to the students that is I have

marks of 10 students. Now, I become benevolent, suddenly and I decide to add 5 marks to each of the students.

So, this is the earlier data set which was 68 had a mean of 59. Now, I am adding 5 marks to each of these students. So, now, the data set becomes $68 + 5$ which is 73; $79 + 5$ which is 84 and so forth $66 + 5$ which is 71. The number of observations remain the same, but I am adding a constant that is I am adding 5 to each value in my data set ok. So, what happens to the mean?

So, earlier my x_1, x_2, x_n were 68, 78, x_3 was 38, my x_{10} was 66. I am going to add y_1 is become 73, $68 + 5$; y_2 is 84 again $78 + 5$; y_3 is 43, $38 + 5$; y_{10} is 71. So, by my definition, if I define what is \bar{x} , \bar{x} is going to be a $\frac{\sum x_i}{10}$; \bar{y} is going to be $\frac{\sum y_i}{10}$. This is what our definitions say.

(Refer Slide Time: 17:35)

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

Statistics for Data Science - I

- └ Numerical summaries
- └ Measures of central tendency

Adding a constant

$$\bar{y} = \frac{y_1 + y_2 + y_3 + \dots + y_n}{n}$$

$$= \frac{x_1 + 5 + x_2 + 5 + x_3 + 5 + \dots + x_{10} + 5}{10}$$

- ▶ Let $y_i = x_i + c$ where c is a constant then $\bar{y} = \bar{x} + c$
- ▶ Example: Recall the marks of students
68, 79, 38, 68, 35, 70, 61, 47, 58, 66.
 - ▶ Suppose the teacher has decided to add 5 marks to each student.
 - ▶ Then the data becomes
73, 84, 43, 73, 40, 75, 66, 52, 63, 71.
 - ▶ The mean of the new data set is $\frac{640}{10} = 64 = 59 + 5$

$$\bar{y} = \bar{x} + 5$$



20 / 60

So, if I add this and I compute y_1 , I find out the y_1 is going to be $\frac{640}{10}, \frac{640}{10}$ which is 64. You can notice that 64 is my \bar{y} . This is my \bar{x} which is 5. So, why does this happen? So, I know that $\frac{y_1+y_2+y_3+\dots+y_{10}}{5 \text{ or } 10}$ is my \bar{y} ; but y_1 is $x_1 + 5$, y_2 is $x_2 + 5$, y_3 is $x_3 + 5$. Similarly, y_{10} is also $x_{10} + 5$. So, you can see that I have $x_1 + x_2 + x_3 + x_{10}$ which makes \bar{y} equal to the following.

(Refer Slide Time: 18:35)

Statistics for Data Science -1
 └ Numerical summaries
 └ Measures of central tendency

Adding a constant

$$\bar{y} = \frac{x_1 + x_2 + x_3 + \dots + x_{10} + 50}{10}$$

$$\bar{y} = \bar{x} + 5$$


- ▶ Let $y_i = x_i + c$ where c is a constant then $\bar{y} = \bar{x} + c$
- ▶ Example: Recall the marks of students
 68, 79, 38, 68, 35, 70, 61, 47, 58, 66.
 - ▶ Suppose the teacher has decided to add 5 marks to each student.
 - ▶ Then the data becomes
 73, 84, 43, 73, 40, 75, 66, 52, 63, 71.
 - ▶ The mean of the new data set is $\frac{640}{10} = 64 = 69 + 5$

So, I have this which is nothing but $\frac{x_1 + x_2 + x_3 + x_{10} + 50}{10}$ because this 5 is added 10 times. Now, $\frac{x_1 + x_2 + x_3 + x_{10}}{10}$ is my \bar{x} that is what we have already seen. $x_1 + x_2 + x_3 + x_{10}$ is my \bar{x} that is something which we have already seen ok.

So, I have this as my $\bar{x} 50 + 10$ is 5, this is the constant. So, I have \bar{y} which is equal to $\bar{x} + c$. So, in summary, when I add a constant to every data point what happens is the mean of the new data set is the old mean plus the same constant.

(Refer Slide Time: 19:42)

B	C	D
DATA	Adding constant	Multiplying with constant
68	73	27.2
79	84	31.6
38	43	15.2
68	73	27.2
35	40	14
70	75	28
61	66	24.4
47	52	18.8
58	63	23.2
66	71	26.4
590	640	236
59	64	23.6
63.5	68.5	25.4
68	73	27.2
210.8888889	210.8888889	33.74222222
10	10	10



So, here what we have done in this is my original data set, I add a constant that is 73 every data set is a constant. This sum is equal to 640 that 640, is what I have here the and the mean is 64 which is equal to $59 + 5$. So, adding a constant to every point in the data set is what we have seen so far.

(Refer Slide Time: 20:12)

Statistics for Data Science -1
└ Numerical summaries
└ Measures of central tendency

Multiplying a constant $\bar{y} = \frac{\sum y_i}{n}$

$$\begin{aligned} &= \frac{\sum c x_i + \sum 73}{n} \\ &= \frac{c \sum x_i + n \cdot 73}{n} \\ &= c \bar{x} + 73 \end{aligned}$$

- ▶ Let $y_i = x_i c$ where c is a constant then $\bar{y} = \bar{x} c$
- ▶ Example: Recall the marks of students
68, 79, 38, 68, 35, 70, 61, 47, 58, 66.
 - ▶ Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.
 - ▶ Then the data becomes
27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4
 - ▶ The mean of the new data set is $\frac{236}{10} = 23.6 = 0.4 \times 59$

29/10

The next thing is what happens to with the mean, when I multiply each observation of the data set with a constant. Again, let us look at what. So, my $y | x_1, x_2, x_n$ is my original data set, I am multiplying it with the constant. Why are we even interested in something of this kind?

For example, I conduct an examination, where the marks are for 100. I conduct an examination for 100; but I want to take only 30% of the grade. So, if the marks when it were for 100 has given me a particular thing, I want to take a particular percentage. So, this c could be a percentage.

So, if recall I go back to the same example here, I decide to scale down each mark by 40%. In other sense, each mark is multiplied by my c here is 0.4. So, you can see that if I multiply each point 68×0.4 is 27.2.

So, that is what I have here. I am multiplying each of my data point with 0.4; 68×0.4 , this is 79×0.4 , this is 66×0.4 ; I multiply each of my data set with c . Now, let us look

at what is the sum of that data set is 236 giving me an average of $\frac{236}{10}$ which is equal to 23.6.

Now, you can notice that again, if I go back what is my \bar{y} my \bar{y} is $\frac{y_1+y_2+\dots+y_{10}}{10}$, but my y_1 is equal to c which times $x_1 + c \times x_2 + c \times x_n$.

This constant is constant for every observation. So, I can remove that constant outside and I can just have $\frac{x_1+x_2+x_3+\dots+x_{10}}{10}$. This is \bar{x} ; hence, \bar{y} is $c \times \bar{x}$ ok. So, you can verify my constant here is 0.4; my \bar{x} was 59, so $c \times 0.4 \times 59$ is what I have is 23.6.

(Refer Slide Time: 23:00)

Section summary



1. Mean or average is a measure of central tendency.
2. Compute sample mean for
 - 2.1 ungrouped data.
 - 2.2 grouped discrete data.
 - 2.3 grouped continuous data.
3. Manipulating data
 - 3.1 Adding a constant to each data point.
 - 3.2 Multiplying each data point with a constant.



Navigation icons: back, forward, search, etc.

So, what we have learned when we studied about mean is we first looked at what is a definition of a mean, mean is one of the most commonly used summaries of data. We defined what was a sample mean and a population mean. But further, most of the course we are going to only deal with sample mean definition.

We saw how to compute the sample mean for discrete ungrouped data and then, we looked at how to compute the sample mean for grouped data, then we saw what would happen when we manipulate data by adding a constant to each data point and by multiplying each data point with a constant. That is what we have seen so far.

(Refer Slide Time: 23:51)

Statistics for Data Science -1
└ Numerical summaries
└ Measures of central tendency

Median

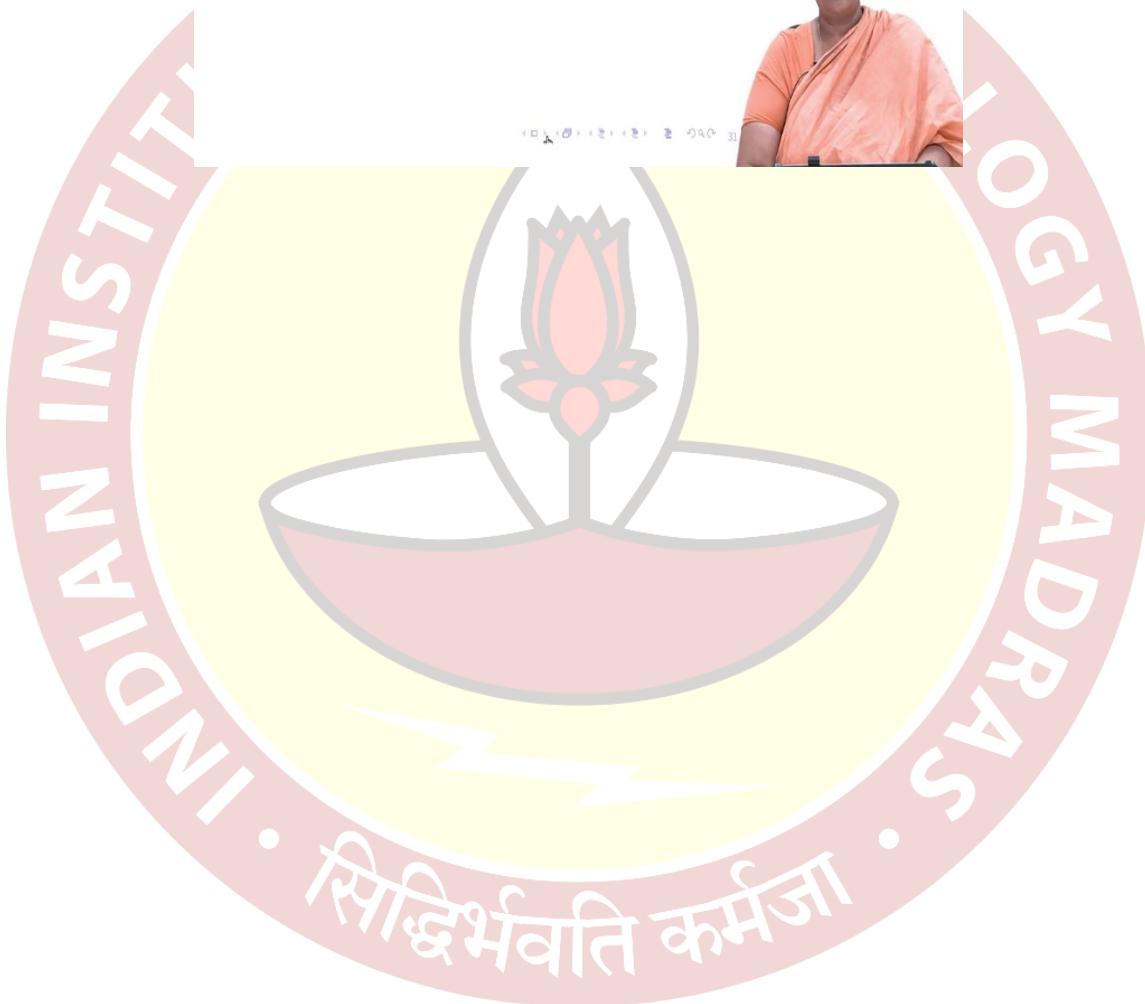


Another frequently used measure of center is the median.
Essentially, the median of a data set is the number that divides the bottom 50% of the data from the top 50%.

Definition

The median of a data set is the middle value in its ordered list.

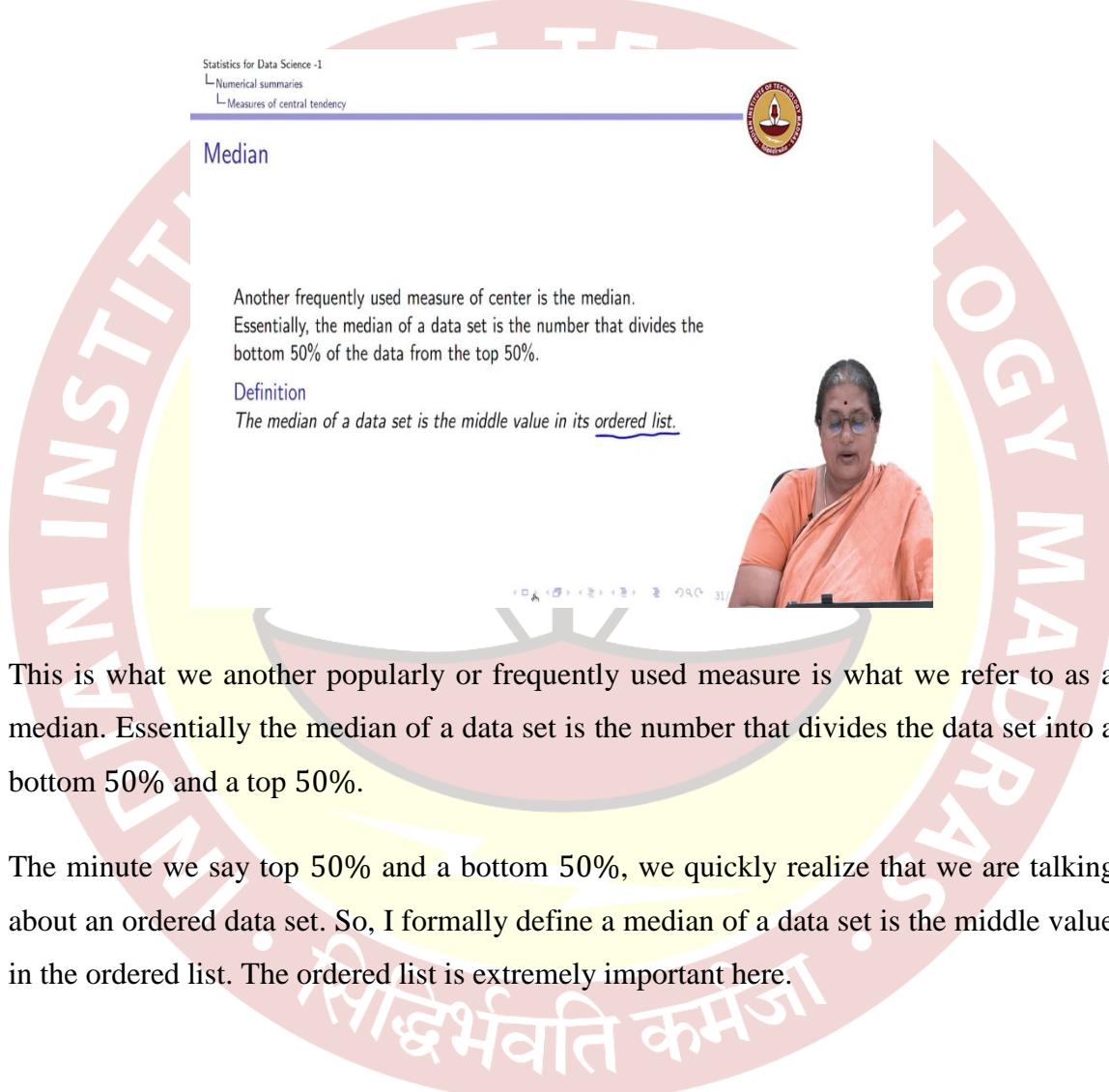
[navigation icons]



Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 3.3
Describing Numerical Data – Median and Mode

(Refer Slide Time: 00:15)



Statistics for Data Science -1
└ Numerical summaries
└ Measures of central tendency

Median

Another frequently used measure of center is the median. Essentially, the median of a data set is the number that divides the bottom 50% of the data from the top 50%.

Definition

The median of a data set is the middle value in its ordered list.

This is what we another popularly or frequently used measure is what we refer to as a median. Essentially the median of a data set is the number that divides the data set into a bottom 50% and a top 50%.

The minute we say top 50% and a bottom 50%, we quickly realize that we are talking about an ordered data set. So, I formally define a median of a data set is the middle value in the ordered list. The ordered list is extremely important here.

(Refer Slide Time: 00:57)

Statistics for Data Science -1
└ Numerical summaries
 └ Measures of central tendency

Steps to obtain median

$$\begin{array}{c} x_1 \ x_2 \ x_3 \ x_4 \ x_5 \ x_6 \\ \downarrow \qquad \qquad \qquad \qquad \qquad \qquad n=6 \\ x_{(5)} + x_{(6)} \\ \hline 2 \end{array}$$



Arrange the data in increasing order. Let n be the total number of observations in the dataset.

1. If the number of observations is odd, then the median is the observation exactly in the middle of the ordered list, i.e. $\frac{n+1}{2}$ observation
2. If the number of observations is even, then the median is the mean of the two middle observations in the ordered list, i.e. mean of $\frac{n}{2}$ and $\frac{n}{2} + 1$ observation



32

So, now how do I compute the median of a data set? The computing data, so I have an ordered list. So, I arrange the data in increasing order. Let n denote the number of observations in the data set. Now, this is important, if the number of observations is odd, then the median of is exactly in the middle of the ordered list.

For example, if my observations $x_1, x_2, x_3, x_4, x_5, 5$; n equal to 5, 5 is odd, then this is assume it is ordered. So, for order data, let me introduce a notation x_1, x_2, x_3, x_4 , and x_5 , x_1 is the order that is the first data, x_2 is the second, x_3 , this is my ordered data, then the data in the $\frac{n+1}{2}, \frac{n+1}{2}$ is $5 + 1, \frac{6}{2}$, third. So, this would be my median. Remember x_1, x_2, x_3, x_4, x_5 are is my data arranged in increasing order. And x_3 which is the third rank data would be my median.

If the number of observations is even, for example, I have x_6 also in this case, my n equal to 6, then the median is going to be your x_3 that is my $(\frac{n}{2})$ observation $+ (\frac{n}{2} + 1)$, x_4 divided by 2, that is what this means. So, if at the median depends on whether the number of observation is an odd number or a even number.

(Refer Slide Time: 03:03)



Example

1. 2, 12, 5, 7, 6, 7, 3
 - 1.1 Arrange the data in increasing order
2, 3, 5, 6, 7, 7, 12
 - 1.2 $n = 7$ odd, median is the $\frac{n+1}{2} = \frac{8}{2} = 4^{\text{th}}$ observation, "6".
2. 2, 105, 5, 7, 6, 7, 3
 - 2.1 Arrange the data in increasing order
2, 3, 5, 6, 7, 7, 105
 - 2.2 $n = 7$ odd, median is the $\frac{n+1}{2} = \frac{8}{2} = 4^{\text{th}}$ observation, "6".
3. 2, 105, 5, 7, 6, 3
 - 3.1 Arrange the data in increasing order
2, 3, 5, 6, 7, 105
 - 3.2 $n = 6$ even, median is the average of $\frac{n}{2}$ and $\frac{n}{2} + 1$ observation
 $= \frac{5+6}{2} = 5.5$.



So, now let us apply this definition and steps to compute the median for the data sets we have already seen before. So, when I have this data 2, 12, 5, 7, 6 and 3, the first step says arrange the data in increasing order. So, the arrangement of data is going to be 2, 3, 2, 3, 5, 6, 7, 7, and 12. I have arranged my data in ascending order.

Now, what is my n in this case? I have 1, 2, 3, 4, 5, 6, 7: n equal to 7 which is odd. So, if n equal to odd, then I apply my, first n equal to odd, the median is $\frac{n+1}{2}$. So, I have n equal to 7 which is odd. So, median is 8 by 2 which is the 4th observation which is equal to 6. So, my median of this data set is 6.

Now, let us look at another example, again the same example. Remember when we are looking at the same examples which we computed the mean for. The difference between the second data set and the first data set is in only one observation which is the second observation here which is 105 for the second data set, and 12 for the first data set.

Remember when we computed the mean, we saw that this one observation actually influenced the mean, and the mean of the first data set and the second data set were very different from each other. Now, let us see what happens to the median of these two data sets.

The number of observations again is the same. I arrange the data in ascending order. So, when I arrange this data in ascending order, I have 2, I have 3, I have 5, 6, 7, 7, and I have

a 105. The number of data is 7 which is again odd. The median is again the 4th observation and which is again 6; it does not change. The median is the *4th* observation which does not change which is equal to 6.

So, what you can immediately notice here is while the mean was very different for these two data sets, the median is the same for both the data sets even though it differs very drastically in an outlier. So, the median is not very sensitive to the outliers the way the mean was.

Now, let us look at the third data set which had only 6 observations. Again I arrange this data in ascending order. So, I have a 2, I have a 3, I have a 5, 6, 7 and 105. Again I have n which is equal to 6. So, my median is going to be the mean of the *3rd* observation and the *4th* observation which is $\frac{5+6}{2}$ which would give me 5.5.

Notice that this 5.5 is not a member of the data set. So, the median need not belong to the data set. Whereas, for the first two data, the date median was a member because we are stay looking at a particular observation; whereas, here I am not looking at a particular observation.

(Refer Slide Time: 06:55)

Statistics for Data Science -1
└ Numerical summaries
 └ Measures of central tendency

Example

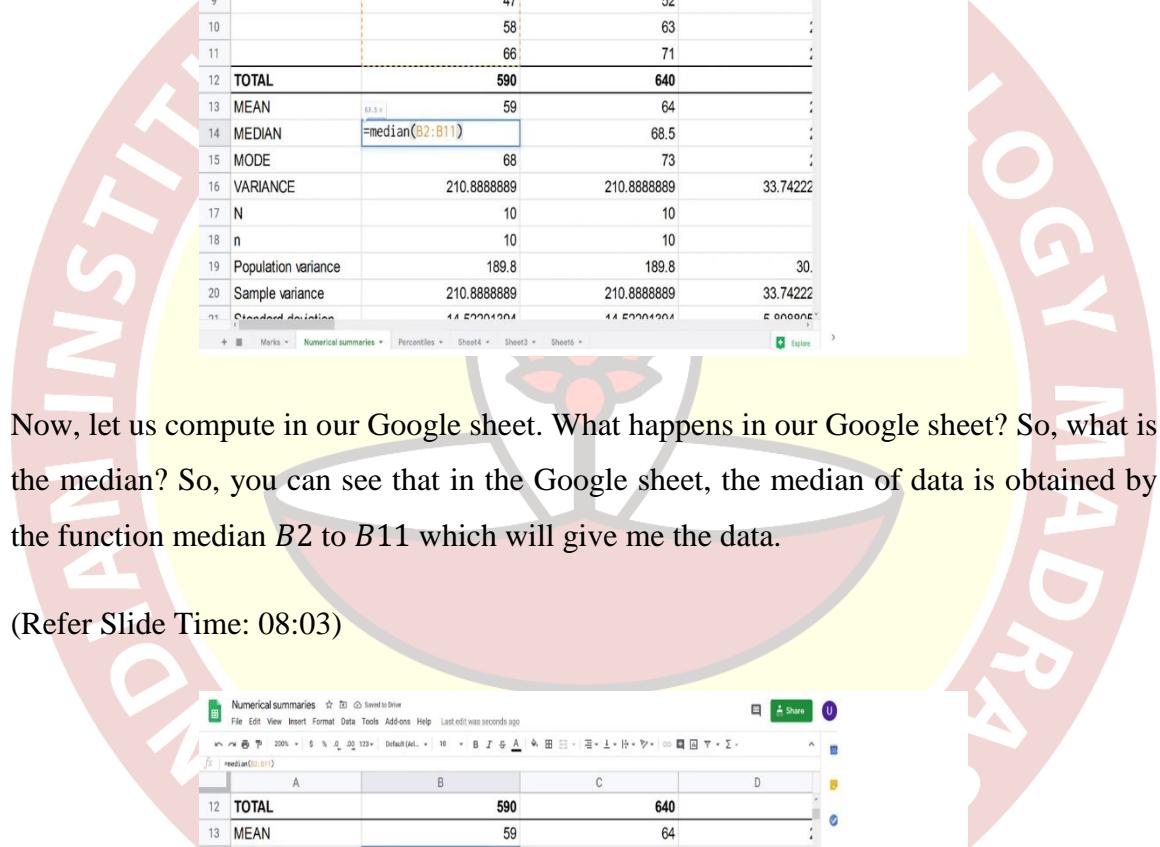


1. 2, 12, 5, 7, 6, 7, 3
 - 1.1 Sample mean = $\frac{2+3+5+6+7+7+12}{7} = 6$
 - 1.2 Sample median = 6
2. 2, 117, 5, 7, 6, 7, 3
 - 2.1 Sample mean = $\frac{2+3+5+6+7+7+117}{7} = 21$
 - 2.2 Sample median = 6

The sample mean is sensitive to outliers, whereas the sample median is not sensitive to outliers.

we already see that there is a significant difference in the mean, but the median remains the same. So, the sample mean is sensitive to outliers, whereas the sample median is not sensitive to outliers.

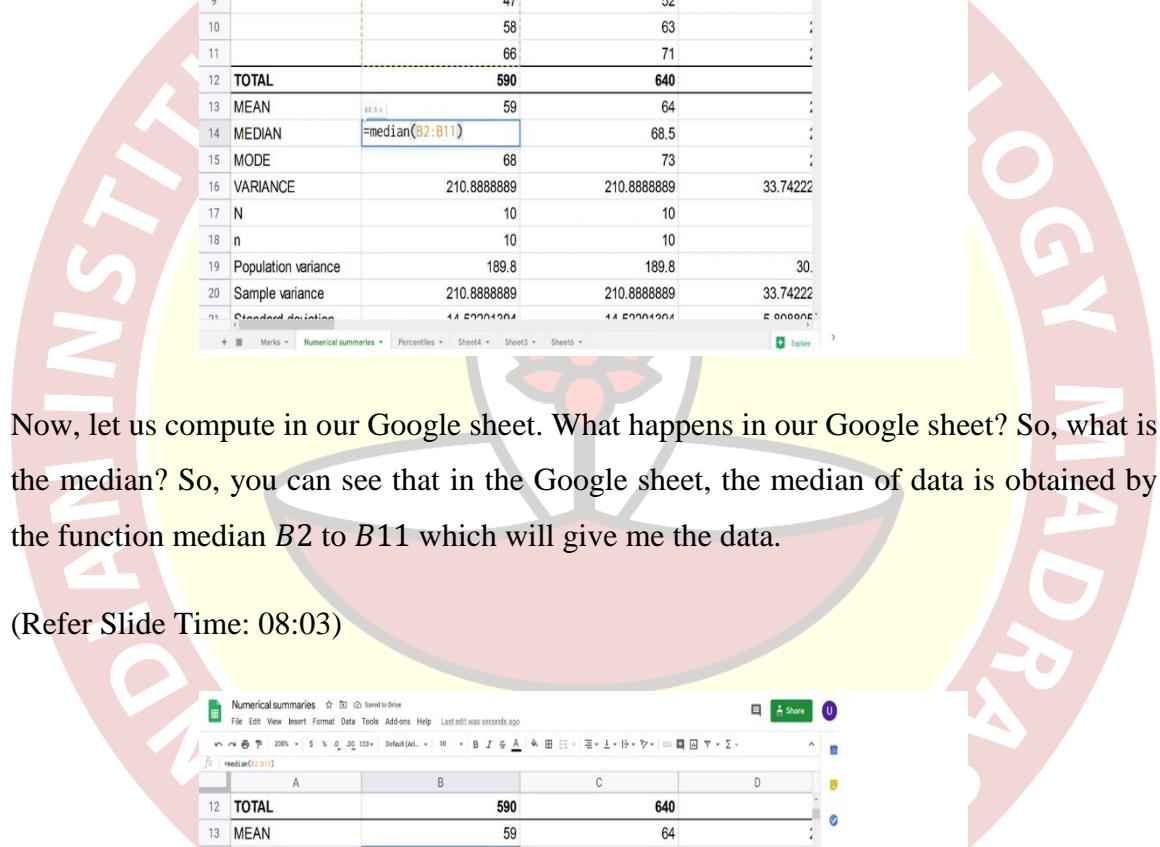
(Refer Slide Time: 07:33)



Numerical summaries			
	B	C	D
5	68	73	
6	35	40	
7	70	75	
8	61	66	
9	47	52	
10	58	63	
11	66	71	
12	TOTAL 590	640	
13	MEAN 59	64	
14	MEDIAN =median(B2:B11)	68.5	
15	MODE 68	73	
16	VARIANCE 210.8888889	210.8888889	33.74222
17	N 10	10	
18	n 10	10	
19	Population variance 189.8	189.8	30.
20	Sample variance 210.8888889	210.8888889	33.74222
21	Standard deviation 14.52201394	14.52201394	5.808805

Now, let us compute in our Google sheet. What happens in our Google sheet? So, what is the median? So, you can see that in the Google sheet, the median of data is obtained by the function median B2 to B11 which will give me the data.

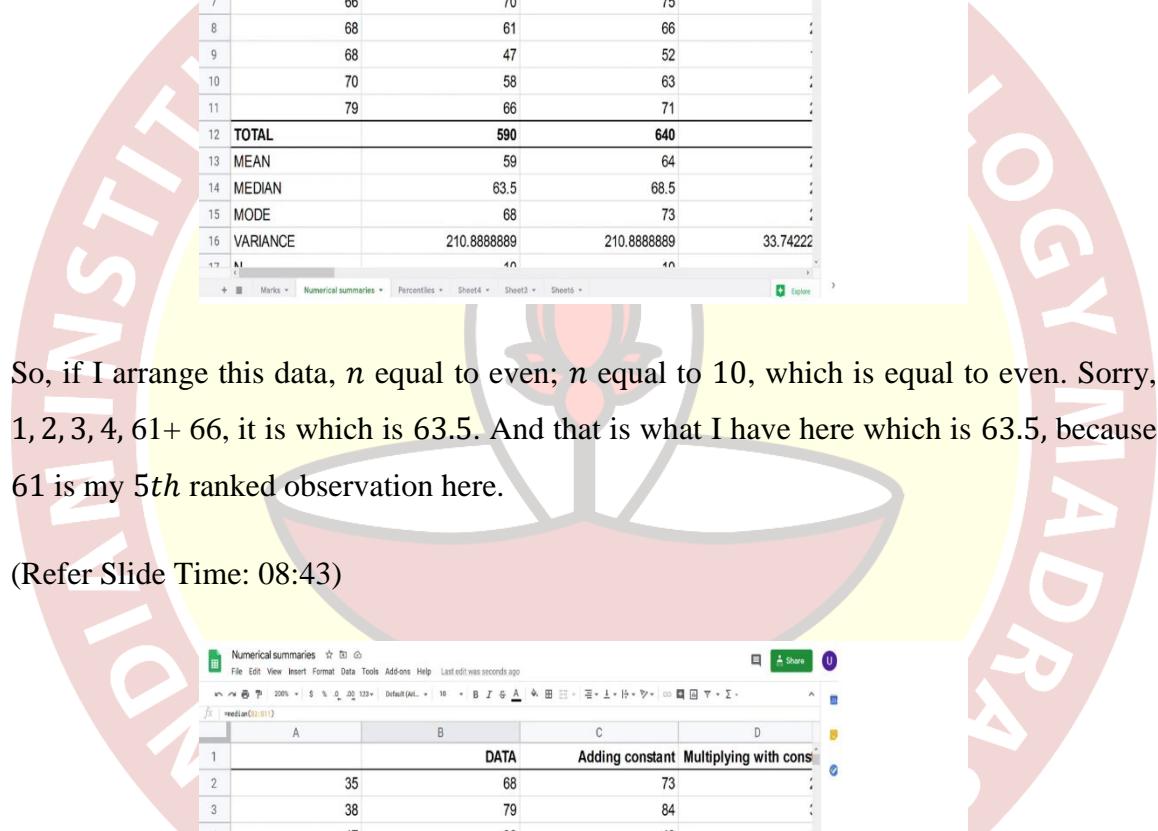
(Refer Slide Time: 08:03)



Numerical summaries			
	B	C	D
12	TOTAL 590	640	
13	MEAN 59	64	
14	MEDIAN 63.5	68.5	
15	MODE 68	73	
16	VARIANCE 210.8888889	210.8888889	33.74222
17	N 10	10	
18	n 10	10	
19	Population variance 189.8	189.8	30.
20	Sample variance 210.8888889	210.8888889	33.74222
21	Standard deviation 14.52201394	14.52201394	5.808805
22		14.52201394	5.808805
23		59	
24		63.5	
25			
26			
27			

So, now I can see that the median here is 63.5. How did we obtain 63.5? I can arrange this data.

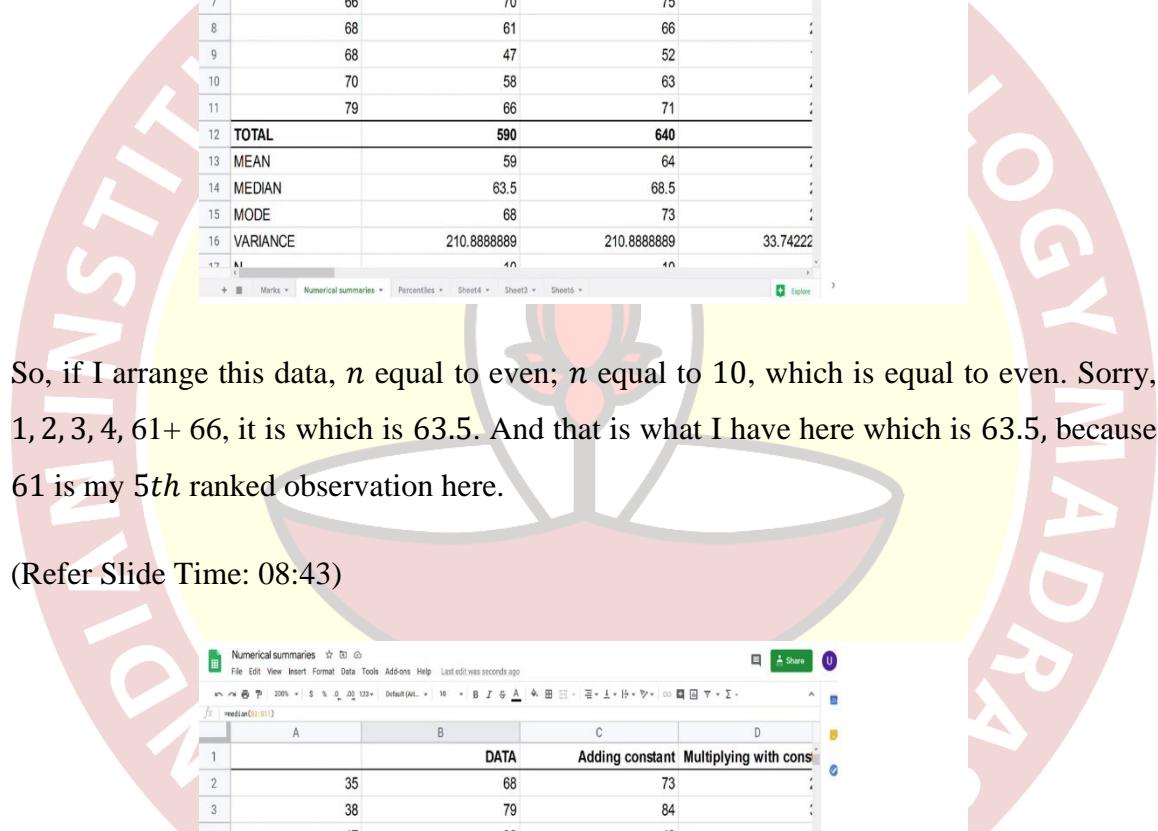
(Refer Slide Time: 08:13)



	A	B	C	D
1		DATA	Adding constant	Multiplying with constant
2	35	68	73	
3	38	79	84	
4	47	38	43	
5	58	68	73	
6	61	35	40	
7	66	70	75	
8	68	61	66	
9	68	47	52	
10	70	58	63	
11	79	66	71	
12	TOTAL	590	640	
13	MEAN	59	64	
14	MEDIAN	63.5	68.5	
15	MODE	68	73	
16	VARIANCE	210.8888889	210.8888889	33.74222

So, if I arrange this data, n equal to even; n equal to 10, which is equal to even. Sorry, 1, 2, 3, 4, 61+ 66, it is which is 63.5. And that is what I have here which is 63.5, because 61 is my 5th ranked observation here.

(Refer Slide Time: 08:43)



	A	B	C	D
1		DATA	Adding constant	Multiplying with constant
2	35	68	73	
3	38	79	84	
4	47	38	43	
5	58	68	73	
6	61	35	40	
7	66	70	75	
8	68	61	66	
9	68	47	52	
10	70	58	63	
11	79	66	71	
12	TOTAL	590	640	
13	MEAN	59	64	
14	MEDIAN	63.5	68.5	
15	MODE	68	73	
16	VARIANCE	210.8888889	210.8888889	33.74222

So, this is 5th $61 + 66$, I have these observations here. So, you can see the $61 + 66$ which is $\frac{137}{2}$ which will give me 63.5 which is the median. I can get this through the command median of the array in Google sheets.

(Refer Slide Time: 09:07)

Statistics for Data Science -1
└ Numerical summaries
└ Measures of central tendency

Adding a constant

- ▶ Let $y_i = x_i + c$ where c is a constant then
new median = old median + c
- ▶ Example: Recall the marks of students
68,79,38,68,35,70,61,47,58,66.
Arranging in ascending order 35,38,47,58,61,66,68,68,70,79
The median for this data is the average of $\frac{n}{2}$ and $\frac{n}{2} + 1$
observation which is $\frac{61+66}{2} = \frac{127}{2} = 63.5$
- ▶ Suppose the teacher has decided to add 5 marks to each student.
- ▶ Then the data in ascending order is
40,43,52,63,66,71,73,73,75,84
- ▶ The median of the new dataset is $\frac{66+71}{2} = \frac{137}{2} = 68.5$
- ▶ Note $68.5 = 63.5 + 5$

So, what happens when we add a constant to the data set? When I add a constant to the data set, again let y_i be x_i plus a constant, where c is a constant, then what happens to my new median? So, let us go back to the example 68, 79, these are the marks of my students. Again I arrange them in ascending order. When I arrange them in ascending order, we saw that the mean is $\frac{61+66}{2}$ which is 63.5.

Again if I decide and the teacher decides to add 5 marks to every student, then the data becomes 40, 43, 47 + 5 which is 52, 63, 66, 73, 75, and 84; I am adding 5 to each point of the data set. Now, again you notice that by adding a constant to the data set does not change the order of the observation.

So, here this $\frac{n}{2}$ observation was 61, and $\frac{n}{2} + 1$ is 66. So, corresponding to 61, I have 66; corresponding to 66 I have 71. The n does not change; the number of observations does not change. So, the median in this case is $\frac{66+71}{2}$, and you can see that it is 68.5. Whereas, 66 is was $61 + 5 - 66$ was 71, 66 + 5. So, the new median is nothing but your old median plus a constant, because the values are the same.

If I have x , so here it is $\frac{x_5+x_6}{2}$ was my old median. My y_5 is $\frac{y_5+y_6}{2}$ is my new median. But y_5 was x_5 plus my constant, y_6 is x_6 plus my constant which is 5. So, I have the new median is $\frac{x_5+x_6}{2}$ which is my old median + 5. So, 5 is the constant.

Old median + 5 is my new median. So, whenever I am adding a constant, the new median is your old median plus the constant. It does not it you are adding that constant to the new median.

(Refer Slide Time: 11:59)

Statistics for Data Science -1
└ Numerical summaries
 └ Measures of central tendency

Multiplying a constant

► Let $y_i = x_i c$ where c is a constant then

$$\text{new median} = \text{old median} \times c = \frac{\frac{y_5+y_6}{2}}{c} = \frac{x_5+c \cdot x_6}{2}$$

What happens when you multiply the entire data set with a constant? When I multiply the data set with an entire constant, again my old data set, so y_1 , so I had $\frac{y_5+y_6}{2}$ which is my new median so, but y_5 is $x_5 + c$, y_6 is going to be $x_6 + c$, I can remove the c . So, I have $\frac{c \times (x_5+x_6)}{2}$, $\frac{x_5+x_6}{2}$, was my old median. So, my old median into the constant will give me the new median. And this we can see in our example.

(Refer Slide Time: 12:47)



Multiplying a constant

- ▶ Let $y_i = x_i c$ where c is a constant then

$$\text{new median} = \text{old median} \times c$$

- ▶ Example: Recall the marks of students
68, 79, 38, 68, 35, 70, 61, 47, 58, 66.
We already know median for this data is 63.5
- ▶ Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.
- ▶ Then the data becomes
27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4
The ascending order is 14, 15.2, 18.8, 23.2, 24.4, 26.4, 27.2, 28, 31.6
The median of new dataset is $\frac{24.4+26.4}{2} = \frac{50.8}{2} = 25.4$
- ▶ Note $25.4 = 0.4 \times 63.5$



So, again recall we know the median is 63.5. If I scale down or each mark is multiplied by 0.4, I saw that this is my what is happening to my data set. Again I arrange the data set in ascending order, my 5th observation here is 24.4, 6th observation is 26.4, the median of the new data set hence is 25.4, which I can verify is 0.4×63.5 ; 25.4 is the new median which is the old median $\times 63.5$.

So, when I go back and see that in my Google sheets, so when I am add a constant, so you can see that the old median is 63.5, when I add a constant of 5, $63.5 + 5$ is 68.5, whereas 63.5×0.4 gives me 25.4. So, this is how we can obtain our new median.

(Refer Slide Time: 14:03)



Mode

Another measure of central tendency is the sample mode.

Definition

The mode of a data set is its most frequently occurring value.



Now, we move on to the third measure of central tendency which we refer to as a mode. We have already seen what is a mode while describing categorical data. We see that the mode as we defined when we talked about categorical data is that observation which has the highest frequency of occurrence.

So, that is the same way we define even for numerical data. So, the mode of a data set as it is given here is the most frequently occurring value, so that is what we refer to as a mode.

(Refer Slide Time: 14:37)



Steps to obtain mode

1. If no value occurs more than once, then the data set has no mode.
2. Else, the value that occurs with the greatest frequency is a mode of the data set.



So, now how do we obtain a mode? Just as we did in the case of the categorical data, even in the numerical data, what we do is we check for the mode by computing or calculating that observation which appears the most number of times. If a value occurs more than once, if no value occurs more than one, the data set has no mode; otherwise that value which occurs with the greatest frequency is the mode of a data set.

(Refer Slide Time: 15:15)

The slide has a navigation bar at the top with 'Data for decision making', 'Numerical summaries', and 'Measures of central tendency'. A circular logo for 'Savitribai Phule Pune University' is on the right. The main content is titled 'Example' and lists three data sets:

1. 2, 12, 5, 7, 6, 7, 3;
7 occurs twice, hence 7 is mode
2. 2, 105, 5, 7, 6, 7, 3
7 is mode
3. 2, 105, 5, 7, 6, 3 no mode

A video player window shows a woman in an orange sari speaking. The video controls include play, pause, and volume buttons.

So, now moving forward we find again we go back to the same data sets which you have considered so far. In this data set I have 2, 12, 5, 7, 7, 6, 3. We can see that 7 appears twice, hence the mode of this data set is 7.

The second data set also 7 occurs twice, again you can see that the difference between the first data set and the second data set is only one observation, namely 12 appears in the first data set, 105 appears in the second data set. The mode again is 7 for this data set.

The third data set has all 6 values that are distinct; hence there is no mode for the third data set. Now, again if you look at the first and second data set, recall when we computed the mean, the mean was very different for both these data sets, the median was the same, the mode is also the same for both the data sets.

(Refer Slide Time: 16:27)

Statistics for Data Science -1
└ Numerical summaries
 └ Measures of central tendency

Adding a constant

► Let $y_i = x_i + c$ where c is a constant then
new mode = old mode + c

$\begin{array}{c} 1, 2, 3 \\ \downarrow \\ 1+c, 2+c, 3+c \end{array}$
 $\begin{array}{c} x_1, x_2, x_3 \rightarrow \text{old mode} \\ \underline{x_2} \\ x_2 + c \end{array}$

Now, as we did in the case of the mean and the median, let us see what happens when we manipulate a data set namely when we add a constant and when we multiply with a constant. When we add a constant to each of the observations of the data set, for example, I have 1, 2, 3, and I am adding a constant to each one of them, this becomes 3, this becomes 4, and this becomes 5, I can see that nothing the characteristic of the data set remains the same.

So, the mode of the data set the new mode is just the old mode $+c$. So, the new mode is that. So, if I have x_1, x_2, x_3 which is my old data set, and suppose x_2 is the mode of my old data set, I add to get y_1, y_2, y_3 which is my x_1 plus a constant, x_2 plus a constant, and x_3 plus a constant. If x_2 is that which appears the most number of time, so it would be $x_1, x_2, x_1, x_2, x_2, x_3$; x_2 is the mode.

So, it becomes y_1, y_2, y_2, y_3 where y_2 appears the most number of times which so you can see that the new mode is $y, x_2 + c$, hence the new mode of my data set is old mode + the constant.

(Refer Slide Time: 18:07)



Adding a constant

- ▶ Let $y_i = x_i + c$ where c is a constant then
 $\underline{\text{new mode}} = \underline{\text{old mode}} + c$
- ▶ Example: Recall the marks of students
68, 79, 38, 68, 35, 70, 61, 47, 58, 66.
The mode for this data is 68
- ▶ Suppose the teacher has decided to add 5 marks to each student.
- ▶ Then the data in ascending order is
40, 43, 52, 63, 66, 71, 73, 73, 75, 84
- ▶ The mode of the new dataset is 73
- ▶ Note $73 = 68 + 5$



Recall again going back to our marks example the data set is 68, 79. I have 10 students, and I can see that the mode for this data set is 68. I add 5 marks just as earlier. And you can see that the data set in ascending order becomes 40, 43, I do not need it in ascending order now, but nevertheless I can see that the mode now is 73 which corresponds to 68 + 5. Hence the new mode is nothing but the old mode + the constant.

(Refer Slide Time: 18:47)



Multiplying a constant

- ▶ Let $y_i = x_i \cdot c$ where c is a constant then

$$\underline{\text{new mode}} = \underline{\text{old mode}} \times \underline{c}$$

$$\begin{array}{l} x_1 x_2 x_3 x_4 \\ \underline{\underline{x_1 x_2 x_3}} \\ y = \underline{\underline{x_1 x_2 x_3}} \\ y = \underline{\underline{x_1 x_2 x_3}} \times c \\ \downarrow \\ \text{mode} \end{array}$$

What happens when we multiply a constant? When we multiply a constant again the new mode is nothing but the old mode times the constant. Again the reasoning is very simple.

Suppose, I have a data set x_1, x_2, x_2, x_3 ; x_2 being the mode here; I have y_1, y_2, y_2, y_3 where y_2 is $x_1 \times$ a constant, y_2 is the mode here. And I know y_2 is nothing but x_2 times the constant; x_2 is the mode for my earlier data set. So, the new mode is old mode \times the constant.

(Refer Slide Time: 19:29)

Statistics for Data Science -1
 └ Numerical summaries
 └ Measures of central tendency

Multiplying a constant

- ▶ Let $y_i = x_i c$ where c is a constant then

$$\text{new mode} = \text{old mode} \times c$$

- ▶ Example: Recall the marks of students
 $68, 79, 38, 68, 35, 70, 61, 47, 58, 66$.
 We already know mode for this data is 68
- ▶ Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.
- ▶ Then the data becomes
 $27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4$
 The mode of new dataset is 27.2
- ▶ Note $27.2 = 0.4 \times 68$



So, again recall the example the mode is 68, this is what we saw from the earlier table example. Now, if the teacher decides to scale down each mark by 40 % and each mark is multiplied by 0.4, the data set becomes the following 27.2, 31.2. The new mode is 27.2, this appears twice. And we can verify that this 27.2 is 0.4 times 68.

(Refer Slide Time: 20:01)

	A	B	C	D	E	F	G	H
1		DATA	Adding constant	Multiplying with constant	Squared Deviations			
2	35	68	73	27.2	81	81	12.96	
3	38	79	84	31.6	400	400	64	
4	47	38	43	15.2	441	441	70.56	
5	58	68	73	27.2	81	81	12.96	
6	61	35	40	14	576	576	92.16	
7	66	70	75	28	121	121	19.36	
8	68	61	66	24.4	4	4	0.64	
9	68	47	52	18.8	144	144	23.04	
10	70	58	63	23.2	1	1	0.16	
11	79	66	71	26.4	49	49	7.84	
12	TOTAL	590	640	236	1898	1898	303.68	
13	MEAN	59	64	23.8				
14	MEDIAN	63.5	68.5	25.4				
15	MODE	68	73	27.2				
16	VARIANCE	210.888889	210.888889	33.7422222				
17	N	10	10	10				
18	n	10	10	10				
19	Population variance	169.8	189.8	30.368				
20	Sample variance	210.888889	210.888889	33.7422222				
21	Standard deviation	14.5201394	14.5201394	5.808805576				
22		14.5201394	14.5201394	5.808805576				
23		59						
24		63.5						

So, if we look at this, we can go back to our numerical summaries. So, you can see that this is my data set. This I arrange my data set in ascending order here. This is the data set. The highlighted portion is the data set in my Google sheets.

So, you can see from this 68 is that value that appears twice, and that is given by the function mode – mode times the data returns the value 68. When I add a constant, the mode of the new data set is 73; 73 is $68 + 5$. And 27.2 is when I multiply it with a constant 27.2 is 68×0.4 , so that is what we have seen.

(Refer Slide Time: 21:05)



Section summary

- ▶ Measures of central tendency
 - 1. Mean
 - 2. Median
 - 3. Mode
- ▶ Impact of adding a constant or multiplying with a constant on the measures.



So, moving forward what we have seen so far is we have studied about the measures of central tendency, namely we looked at what is the mean, we define both the population mean and the sample mean.

But then our discussion centered mostly around the sample mean. Then we moved on to define what is a median of a data set, and then what is a mode of a data set. For each one of these operations or measures, we saw what was the impact of adding a constant or multiplying with a constant on the measures.

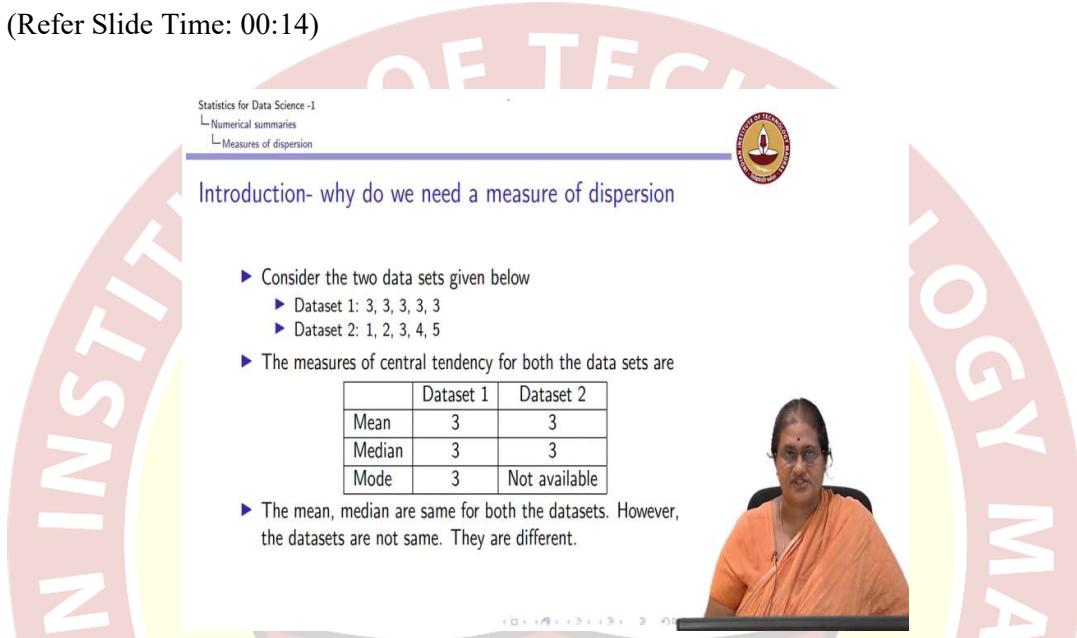
That is what we have.



Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 3.4
Describing Numerical Data - Measures of dispersion: Range, Variance, and Standard deviation

(Refer Slide Time: 00:14)



Statistics for Data Science - 1
└ Numerical summaries
└ Measures of dispersion

Introduction- why do we need a measure of dispersion

▶ Consider the two data sets given below

- ▶ Dataset 1: 3, 3, 3, 3, 3
- ▶ Dataset 2: 1, 2, 3, 4, 5

▶ The measures of central tendency for both the data sets are

	Dataset 1	Dataset 2
Mean	3	3
Median	3	3
Mode	3	Not available

▶ The mean, median are same for both the datasets. However, the datasets are not same. They are different.



The next thing which we are going to look at is Measures of Dispersion. So, why do we need to know about measures of dispersion? So, measures of central tendency actually captures what we call the center or the typicalness of the dataset. So, what is measure of dispersion capture? Why first of all even before going to define what is a measure of dispersion, let us understand why do we need a measure of dispersion.

Towards this, let us look at two datasets. The 1st dataset is or both the datasets have 5 observations each, the 1st dataset has observations 3, 3, 3, 3, and 3. The 2nd dataset has observations 1, 2, 3, 4, 5.

Let us work out the measures of central tendency for this dataset. So, if we work out the measures of central tendency for this dataset, we observe the following: the mean of the 1st dataset is 3, the mean of the 2nd dataset is also 3 because the mean of the 1st dataset is $3 + 3 + 3 + 3 + 3$ which is 15 divided by 5 which is 3. The 2nd dataset is $1 + 2 + 3 + 4 + 5$ which is again another 15 divided by 5 which is equal to 3.

We look at the median, the number of observations is odd which is 5 so, $5 + 1$ is 6, 6 by 2 the third observation, the third observation and dataset 1 is 3 again the median for the second observation also is 3. The mode, the mode for the 1st dataset 3 appears there is only one value and 3 which appears 5 times so, the mode is 3 whereas, for the 2nd dataset, there is no mode.

However, when you look at this dataset and only if the numerical summaries and the measures of central tendency are given to you, you somehow tend to believe that both the datasets are very similar in nature because the mean and the median of both these datasets are the same. However, we see that both these datasets are very different from each other.

(Refer Slide Time: 02:52)

Statistics for Data Science -1
└ Numerical summaries
└ Measures of dispersion

Measures of dispersion

- ▶ To describe that difference quantitatively, we use a descriptive measure that indicates the amount of variation, or spread, in a data set.
- ▶ Such descriptive measures are referred to as
 - ▶ measures of dispersion, or
 - ▶ measures of variation, or
 - ▶ measures of spread.
- ▶ In this course we will be discussing about the following measures of dispersion.
 1. Range.
 2. Variance.
 3. Standard deviation.
 4. Interquartile range.

A photograph of a woman in an orange sari speaking at a podium is visible in the bottom right corner of the slide.

So, they when they are different, we want to see that is there any other measure that can capture this difference and hence, we need to understand what is a measure of spread or dispersion. The first understanding we need to have is I have to describe this difference quantitatively which will actually tell me what is the amount of variation, what is the amount of spread of a dataset.

So, the descriptive measures are popularly referred to as measures of dispersion or measures of variance or measures of spread. What are the key measures of variation we are going to understand or dispersion we are going to understand in this course?

We start with defining what is a range, then we go on to define what is variance, once we establish what is a variance, variance is the most frequently used measure of dispersion, we again we define what is a standard deviation and then, after introducing what are percentiles, we will introduce a notion of a interquartile range.

(Refer Slide Time: 03:55)

Range

Definition
The range of a data set is the difference between its largest and smallest values.

- The range of a data set is given by the formula

$$\text{Range} = \text{Max} - \text{Min}$$

where Max and Min denote the maximum and minimum observations, respectively.

	Dataset 1	Dataset 2
3,3,3,3,3	1,2,3,4,5	
Max	3	5
Min	3	1
Range	0	4

So, let us go ahead and understand what is a range. A range is defined range of a dataset is defined as the difference between its largest and smallest value. So, the range as the name suggest is a difference between the largest and the smallest value. So, the range is basically maximum - minimum where maximum is the largest and minimum is the smallest value.

So, let us go back to the two datasets we have. The 1st dataset was 3, 3, 3, 3, 3, the 2nd dataset was 1, 2, 3, 4, 5 the maximum of the 1st dataset is 3 because I have only one data value in that, the maximum of the 2nd dataset is 5, the minimum of the 1st dataset is again 3, the minimum of the 2nd dataset is 1. Hence, the range of my 1st dataset is 0 whereas; the range of my 2nd dataset is 4.

(Refer Slide Time: 05:12)



Range sensitive to outliers

- ▶ Range is sensitive to outliers. For example consider two datasets as given below

	Dataset 1	Dataset 2
	1,2,3,4,5	1,2,3,4,15
Max	5	15
Min	1	1
Range	4	14

- ▶ Though the two datasets differ only in one datapoint, we can see that this contributes to the value of Range significantly. This happens because the range takes into consideration only the Min and Max of the dataset.



So, we can see that the range in a sense tells us more about the datasets, than what the measures of central tendency actually revealed. But; however, what is the problem with the range? Now, again consider the two datasets. The 1st dataset is just 1, 2, 3, 4, 5 and the 2nd dataset is 1, 2, 3, 4 and 15. These two datasets differ with each other only in one observation namely I have a 5 here whereas, I have a 15 here. So, this is the key thing where they are different.

So, the maximum of the 1st dataset is a 5, the maximum of the 2nd dataset is a 15, the minimum for both the datasets is 1. So, the range of the 1st dataset is $5 - 1$ which is a 4 whereas, the range of the 2nd dataset is 14. So, as you can see very similar to what we observed when we discussed about the mean, we see that the range is extremely sensitive to outliers. So, the range is an extremely sensitive measure because the range takes into consideration only the extreme values to compute it the formula.

(Refer Slide Time: 06:20)

Variance

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

1	2	3	4	5
-2	-1	0	1	2
1	2	3	4	5
3	0	1	2	3

- ▶ In contrast to the Range, the variance takes into account all the observations.
- ▶ One way of measuring the variability of a data set is to consider the deviations of the data values from a central value

Now, the next measure of dispersion which we are going to talk about and this is the most frequently used measure of dispersion is what we refer to as the variance of a dataset. In contrast to the range, the variance takes into account all the observations. What do we mean by this? Again the range takes into account only the minimum and the maximum namely the extreme observations are taken into consideration when you actually compute the range whereas, the variance takes into account all the observations.

So, the one way of measuring the variability is to consider the deviations of the data value from a central value. What do we mean by this? Suppose, I have a data x_1, x_2, x_3, x_4, x_5 , I have a measure of central tendency for now let me call that \bar{x} , I have this measure of central tendency which I have defined. So, I have a measure of central tendency which I have defined as \bar{x} , okay.

So, the deviation of the data values from this central value is $x_1 - \bar{x}, x_2 - \bar{x}$ and so for $x_n - \bar{x}$ this is what I mean by the deviation or the difference of each of the data points from its central value and the central value I have chosen here is the mean, okay.

Now, one thing is in a given dataset for example, let us again take a 1, 2, 3, 4, 5 I know \bar{x} for this dataset is 3, the deviation of 1 from 3 is -2, 2 from 3 is -1, 3 from itself is 0, 4 from 3 is +1 and 5 from 3 is 2. We can see that these are the deviations of the dataset, but then after so, I need an aggregate measure, I just cannot give the deviation. One possibility is if I sum up all the deviations, I see it goes to 0. Hence, again I see that

summing up the deviations is not a very good measure of the variability even though I have taken all the data points into consideration.

(Refer Slide Time: 08:59)

Statistics for Data Science - 1

- └ Numerical summaries
- └ Measures of dispersion

Population variance and sample variance

Recall when we refer to a dataset from a population, we assume the dataset has N observations, whereas, when refer to a dataset from a sample, we assume the dataset has n observations.

- The variance is computed using the following formulae

► Population variance: $\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_N - \mu)^2}{N}$

$$\sigma^2 = \frac{x_1 (x_1 - \mu)^2 + x_2 (x_2 - \mu)^2 + \dots + x_N (x_N - \mu)^2}{N}$$

A woman in an orange sari is speaking in front of a red banner with white text.

So, what variance the population variance and the sample variance does is the following. It look at the deviation of each of my dataset from data point, from the central value and it squares it up. It adds the squares of the deviation. So, sum of squared deviations from the central value and it averages it. Again I repeat, it takes the deviation of every data's point from its central value, it squares the deviation and adds up all the deviations and divides it by a number.

Now, if it we are talking about a population variance, then the variance of a population variance or the population variance given by σ^2 remember population mean is μ so, I have x_1 , I have x_2 , I have x_N which are my population units.

$x_1 - \mu$ is the deviation of the first unit from the mean, $x_2 - \mu$ is the deviation of the second unit from the mean, $x_N - \mu$ is the deviation of the nth unit from the mean I square each of these deviations, I add them up, I get the numerator. If I divide the total sum of square deviations with the total number of observations, I refer it to the population variance.

(Refer Slide Time: 10:57)



Population variance and sample variance

Recall when we refer to a dataset from a population, we assume the dataset has N observations, whereas, when refer to a dataset from a sample, we assume the dataset has n observations.

- The variance is computed using the following formulae

$$\begin{aligned} \text{Population variance: } \sigma^2 &= \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \dots + (x_N-\mu)^2}{N} \\ \text{Sample variance: } s^2 &= \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \dots + (x_n-\bar{x})^2}{n-1} \end{aligned}$$

$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$



However, when I talk about the sample variance again I have x_1, x_2, \dots, x_n , n is my sample size, the mean is given by \bar{x} . I look at the deviations of each data point from the mean and I square it up, the sum of the square deviations divided by $n - 1$. Notice the difference between the population variance and the sample variances.

The population variance I divide the sum of square deviations by the total number of observations. When I refer to the sample variance, I divide the total number of sample squared deviations by the number of total observations - 1. There is a reason to do so. The explanation is out of the scope of this particular course and you will learn about this as you go forward.

The notation for a population variance I refer as σ^2 , sample variance I refer as s^2 , okay. So, I repeat the numerator be it the population variance or the sample variance is the sum of squared deviations of a data point from its mean value.

(Refer Slide Time: 12:25)



Population variance and sample variance

Recall when we refer to a dataset from a population, we assume the dataset has N observations, whereas, when refer to a dataset from a sample, we assume the dataset has n observations.

- ▶ The variance is computed using the following formulae
 - ▶ Population variance: $\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_N - \mu)^2}{N}$
 - ▶ Sample variance: $s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$
- ▶ The numerator is the sum of squared deviations of every observation from its mean.
- ▶ The denominator for computing population variance is N , the total number of observations.
- ▶ The denominator for computing sample variance is $(n - 1)$.
The reason for this will be clear in forthcoming courses on statistics.



So, the numerator is a sum of square deviation, the denominator for population variance is N whereas, it is $n - 1$. The reason will become very clear in the forthcoming courses, but for now we are going to restrict ourselves to sample variance.

(Refer Slide Time: 12:43)



Example

- ▶ Recall marks of students obtained by ten students in an exam is 68, 79, 38, 68, 35, 70, 61, 47, 58, 66
- ▶ The mean was computed to be 59.
- ▶ The deviations of each data point from its mean is given in the table below:



So, again go back to looking at the same example that is the marks obtained by the 10 students in an exam. So, the mean remember, we computed the mean to be 59. So, let us compute this deviations.

(Refer Slide Time: 13:03)



	Data	Deviation from mean $(x_i - \bar{x})$	Squared deviations $(x_i - \bar{x})^2$
1	68	9	81
2	79	20	400
3	38	-21	
4	68	9	81
5	35	-24	
6	70	11	
7	61	2	4
8	47	-12	144
9	58	-1	1
10	66	-7	49
Total	590	0	



So, you can see that the deviation is given here $68 - 59$ is 9, $79 - 59$ is 20, $38 - 59$ is -21 and the square deviations here I have a 81, I have a 400.

(Refer Slide Time: 13:27)



	Data	Deviation from mean $(x_i - \bar{x})$	Squared deviations $(x_i - \bar{x})^2$
1	68	$68 - 59$	81
2	79	$79 - 59$	400
3	38	$38 - 59$	441
4	68	$68 - 59$	81
5	35	$35 - 59$	576
6	70	11	121
7	61	2	4
8	47	-12	144
9	58	-1	1
10	66	$66 - 59$	49
Total	590	0	1898

$$\sum (x_i - \bar{x})^2 = 1898$$

$n = 10$



So, fourth I keep writing this and I can see that I can find out what are the sum of the squared deviations for each one of the dataset. So, this is $68 - 59$, this is $79 - 59$, this is $38 - 59$, $68 - 59$, $35 - 59$. So, you can see that and this is $66 - 59$, this is 7^2 , 1^2 , 11^2 , 24^2 , the sum of the square deviation -24^2 , -1^2 , -7^2 . So, the numerator $\sum(x_i - \bar{x})^2$ is 1898. I have n equal to 10 to compute the sample variance I divided by $10 - 1$ which is 9.

(Refer Slide Time: 14:26)

Statistics for Data Science -1
└ Numerical summaries
└ Measures of dispersion



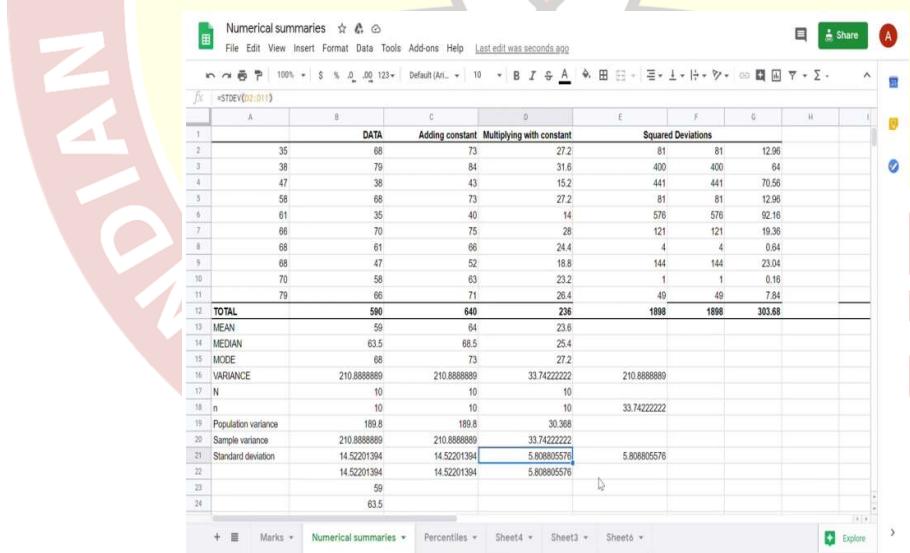
	Data	Deviation from mean $(x_i - \bar{x})$	Squared deviations $(x_i - \bar{x})^2$
1	68	9	81
2	79	20	400
3	38	-21	441
4	68	9	81
5	35	-24	576
6	70	11	121
7	61	2	4
8	47	-12	144
9	58	-1	1
10	66	-7	49
Total	590	0	1898

1. Population variance = $\frac{1898}{10} = 189.8$
 2. Sample variance = $\frac{1898}{9} = 210.88$



And you can see that, my population variance is 189.8 whereas, my sample variance is 210.88, okay.

(Refer Slide Time: 14:42)



The screenshot shows a Google Sheets document with the following data and formulas:

	A	B	C	D	E	F	G	H
1		DATA	Adding constant	Multiplying with constant	Squared Deviations			
2	35	68	73	27.2	81	81	12.96	
3	38	79	84	31.6	400	400	84	
4	47	38	43	15.2	441	441	70.56	
5	58	68	73	27.2	81	81	12.96	
6	61	35	40	14	576	576	92.16	
7	66	70	75	28	121	121	19.36	
8	68	61	66	24.4	4	4	0.64	
9	68	47	52	18.8	144	144	23.04	
10	70	58	63	23.2	1	1	0.16	
11	79	66	71	26.4	49	49	7.84	
12	TOTAL	590	640	236	1898	1898	303.68	
13	MEAN	59	64	23.6				
14	MEDIAN	63.5	68.5	25.4				
15	MODE	68	73	27.2				
16	VARIANCE	210.8888889	210.8888889	33.74222222	210.8888889			
17	N	10	10	10				
18	n	10	10	10	33.74222222			
19	Population variance	189.8	189.8	30.368				
20	Sample variance	210.8888889	210.8888889	33.74222222				
21	Standard deviation	14.52201394	14.52201394	5.808805576	5.808805576			
22		14.52201394	14.52201394	5.808805576				
23		59						
24		63.5						

Formulas used in the sheet:

- =STDEV(B2:B12)
- =STDEV.S(B2:B12)
- =STDEV.P(B2:B12)

So, now let us look at how to compute this using our Google sheet. So, in our Google sheets, this is again my data is what I have highlighted here, okay. So, now, if you look at this portion here, this is the squared deviation. So, you can see that this square deviation is B2, B2 is my data point, B13 is my mean whole square. So, this corresponds

to the 81 I have here okay. Similarly, I have a 400 the second data point, I have 441 the third and so forth my total sum of square deviations as highlighted here is 1898.

So, this divided by my 10 would be 189.8, but when I am dividing it by 9 I get 210.88. Now, the same thing if you use the function VAR.S; VAR.S, S to represent the sample statistic VAR.S of the array returns the sample variance. I reply I repeat VAR.S returns the sample variance and we can see that this is equal to the sum of the square deviation divided by 9 and I get the same value.

Now, the population variance is nothing but now as in the earlier case, let us see what happens to these variance when we manipulate the dataset. What we mean by this is what would happen to the dataset if I add a constant or I multiply each one of the values with the constant.

(Refer Slide Time: 17:10)

Statistics for Data Science -1
└ Numerical summaries
└ Measures of dispersion

Adding a constant

► Let $y_i = x_i + c$ where c is a constant then
new variance = old variance

For new dataset

$$\sum_{i=1}^n (y_i - \bar{y})^2$$

$$[x_i + c - (\bar{x} + c)] \rightarrow x_i + c - \bar{x} - c = (x_i - \bar{x})$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$V(y) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

$$= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$V(y)$$

A photograph of a woman in an orange sari sitting at a desk, looking towards the camera. She is part of the video recording for the slide.

So, what would happen if I add a constant? So, I have y_i is $x_i + c$ where c is a constant. So, again let me just look at 3 numbers y_1, y_2, y_3 . I have x_1, x_2, x_3 . \bar{x} is the mean of these three numbers, \bar{y} is the mean of these three numbers. We have already seen \bar{y} is $c + \bar{x}$ this is already what we have seen.

So, now, when I compute my variance for the new dataset, I get a $\sum_{i=1}^n (y_i - \bar{y})^2$. My y_i so, it is $x_i + c$ that is each y_i , \bar{y} is $\bar{x} + c$ so, you can see that this is $x_i + c - \bar{x} - c$ which

is same as $x_i - \bar{x}$, is that clear. So, I have $y_i - \bar{y}$ is $x_i + c$ because each y_i is $x_i + c$, \bar{y} is $\bar{x} + c$. So, it is $x_i + c - \bar{x} - c$ which is these two get cancelled out and I get $x_i - \bar{x}$.

So, I have the numerator $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (x_i - \bar{x})^2$. Hence, $v(y) = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = v(x)$.

Hence, if I add a constant to every value of my dataset, the variance of the dataset does not change the new variance is equal to the old variance and that is what we have just seen.

(Refer Slide Time: 19:49)

Statistics for Data Science -1

- └ Numerical summaries
- └ Measures of dispersion

Adding a constant

- ▶ Let $y_i = x_i + c$ where c is a constant then
new variance = old variance
- ▶ Example: Recall the marks of students
68,79,38,68,35,70,61,47,58,66. has sample variance 210.88
- ▶ Suppose the teacher has decided to add 5 marks to each student.
- ▶ Then the data is
73, 84, 43, 73, 40, 75, 66, 52, 63, 71
- ▶ The variance of the new dataset is $\frac{1898}{9} = 210.88$ ✓
- ▶ In general, adding a constant does not change variability of a dataset, and hence it is the same.

So, let us look at a example to see what is happening. Recall we just computed the marks of the students and the sample variance is 210.88, we add find marks the new dataset is this and you can see that the new variance of this dataset is also 210.88. In general, we have adding a constant does not change the variability of a dataset.

Let us go back to this example. Again this is my original data the one that is highlighted, the variance was 210.88, I add a constant and I get this as my dataset and see that the variance for both these datasets is the same.

(Refer Slide Time: 20:58)

Statistics for Data Science - I
└ Numerical summaries
└ Measures of dispersion

Multiplying a constant

$$\begin{aligned}
 y_1 &= x_1 c \\
 y_2 &= x_2 c \\
 y_n &= x_n c \\
 \bar{y} &= \bar{x}c \\
 (y_i - \bar{y})^2 &= (x_i c - \bar{x}c)^2 = c^2(x_i - \bar{x})^2 \\
 v(y) &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{c^2}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = c^2 v(x)
 \end{aligned}$$

So, adding a constant does not change variability of a dataset. What happens when we multiply the dataset with a constant? Again, I have my $y_1 = x_1 c$, $y_2 = x_2 c$, $y_n = x_n c$, I know $\bar{y} = \bar{x}c$. So, my $y_i - \bar{y} = x_i c - \bar{x}c = c(x_i - \bar{x})$, okay.

I repeat, for every $y_i - \bar{y} = c(x_i - \bar{x})$. So, $(y_i - \bar{y})^2 = c^2(x_i - \bar{x})^2$. So, $\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = c^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$. Hence, $v(y) = c^2 v(x)$, is that clear. So, this is my dataset I already know $\bar{y} = \bar{x}c$. I just substitute the values and I get the $v(y) = c^2 v(x)$.

(Refer Slide Time: 22:53)

Statistics for Data Science - I
└ Numerical summaries
└ Measures of dispersion

Multiplying a constant

- ▶ Let $y_i = x_i c$ where c is a constant then

$$\text{new variance} = c^2 \times \text{old variance}$$

- ▶ Example: Recall the marks of students 68, 79, 38, 68, 35, 70, 61, 47, 58, 66.
We already know variance for this data is 210.88
- ▶ Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.
- ▶ Then the data becomes 27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4
The mean of new dataset is 23.6
- ▶ The sum of squared deviations from mean = 303.68 and the variance = $\frac{303.68}{9} = 33.74$. We can verify that $33.74 = 0.4^2 \times 210.88$.

So, we have the following that if $y_i = x_i c$, the new variance = $c^2 \times$ old variance, we have just established that relation. We can verify that using our dataset. I already know the data the variance for the dataset is 210.88, I multiply it with 0.4, I know this is my dataset the mean of this new dataset is 23.6 and I can compute that the variance which is 33.74 is 0.4 square 210.88.

So, you can see that when I multiply with a constant, I get 33.74 is my variance and I can verify that this is 0.4 times the old variance. So, this 0.4 times the old variance is my new variance. So, this is how we compute the dataset with adding a constant and multiplying with a constant.

(Refer Slide Time: 24:08)

Statistics for Data Science - I
└ Numerical summaries
└ Measures of dispersion

Standard deviation

- ▶ Another very useful measure of dispersion is the standard deviation.

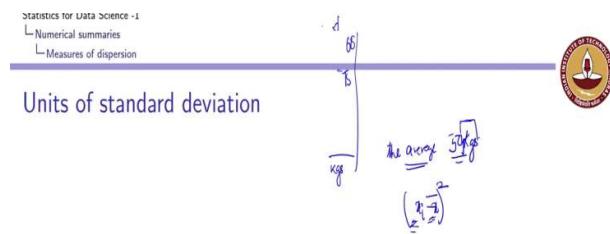
Definition
The quantity

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}}$$

which is the square root of sample variance is the sample standard deviation.

Another useful measure of dispersion is the standard deviation. The standard deviation is nothing, but the square root of the variance. Now, why are we interested in the standard deviation? The standard deviation is referred to by the small lower-case letter s . So, the standard deviation is the square root of the sample variance. Similarly, I can define this population standard deviation also.

(Refer Slide Time: 24:40)



22/3



So, why do I require a standard deviation? Remember whenever I talk about a numerical measure, there are units associated with the numerical measures. For example, if I were looking at ages of people or I was looking at heights of students or I was looking at the weights of students 68, suppose 75 instead of marks they were weights of students measured in kilograms, then the average which was 59 which I did not give any units, if it were weights would have been 59 kilograms.

In other words, the average of a dataset has units which is same as the original measurement. I repeat, the average of a dataset have the same units as that of the original dataset. But when I am computing the variance, what I do is I will take the dataset. So, average has the same units as the dataset.

So, if it is kilogram this is a kilogram, this is a kilogram I am squaring it up so, difference will also be kilogram so, the square is not going to be kilogram, it would be kilogram square I add units of kilogram square so, my units of the variance is square of the units of the original variable.

(Refer Slide Time: 26:23)



Units of standard deviation

- ▶ The sample variance is expressed in units of square units if original variable. For example, instead of marks if the data were weights of 10 students measured in kilograms. Then the unit of variance would be $(\text{kilogram})^2$
- ▶ The sample standard deviation is measured in the same units as the original data. That is, for instance, if the data are in kilograms, then the units of standard deviation are also in kilograms.



So, to overcome this, what we do is we define what is called the standard deviation so, that the units of measurement of the standard deviation and the original units of measurement are the same. So, that if I am looking at variance, I have a unit which is $(\text{kilogram})^2$, I bring it back to kilograms. So, the sample standard deviation is measured in the same units as the original data. If the data is in kilograms, the units are also in kilograms.

(Refer Slide Time: 27:03)



Adding a constant

$$\begin{aligned}x_1 &= x_1 \\x_2 &= x_2 \\&\vdots \\x_n &= x_n \\y_1 &= y_1 \\y_2 &= y_2 \\&\vdots \\y_n &= y_n \\y_i &= x_i + c \\f(y_i) &= f(x_i) \\f(y_i) &= \sqrt{f(x_i)}\end{aligned}$$



So, what happens when we add a constant to the data? So, again I have x_1, x_2, \dots, x_n , I am getting y_1, y_2, \dots, y_n where each $y_i = x_i + c$. I know $v(y_i) = v(x_i)$. Hence, $\sqrt{v(y_i)} = \sqrt{v(x_i)}$.

(Refer Slide Time: 27:37)

Adding a constant

- ▶ Let $y_i = x_i + c$ where c is a constant then
new variance = old variance
- ▶ Example: Recall the marks of students
68,79,38,68,35,70,61,47,58,66. has sample variance 210.88
- ▶ Suppose the teacher has decided to add 5 marks to each student.
- ▶ Then the data is
73, 84, 43, 73, 40, 75, 66, 52, 63, 71
- ▶ The variance of the new dataset is $\frac{1898}{9} = 210.88$
- ▶ the standard deviation of the new dataset is
 $\sqrt{210.88} = 14.522$
- ▶ In general, adding a constant does not change variability of a dataset, and hence it is the same.

new var = old var
new standard deviation = old standard deviation

INSTITUTE OF TECHNOLOGY MANGALURU
सिद्धिर्भवति कर्मजा

And hence, the standard deviation is also the same, the new variance is equal to old variance. Recall this, the sample variance was the same, the sample standard deviation is also the same. Hence, the constant does not change the variability. Both new variance is equal to old variance, new standard deviation is equal to old standard deviation. In other words, adding a constant does not change the variability of a dataset.

(Refer Slide Time: 28:25)

Statistics for Data Science -1
└ Numerical summaries
└ Measures of dispersion

Multiplying a constant $y_i = c x_i$

$$\begin{aligned} v(y) &= c^2 v(x) \\ s(y) &= \sqrt{v(y)} = \sqrt{c^2 v(x)} \\ &= c \sqrt{v(x)} \\ &= c s(x) \end{aligned}$$

What happens when we multiply a constant? Again, you know that if $y_i = cx_i$, we saw the $v(y) = c^2 v(x)$. Hence, $SD(y) = \sqrt{v(y)} = \sqrt{c^2 v(x)} = c \sqrt{v(x)} = c SD(x)$.

(Refer Slide Time: 28:56)

Statistics for Data Science -1
└ Numerical summaries
└ Measures of dispersion

Multiplying a constant

- ▶ Let $y_i = x_i/c$ where c is a constant then $\frac{\text{new std dev}}{\text{old std dev}} = \frac{1}{c}$
- ▶ Example: Recall the marks of students 68,79,38,68,35,70,61,47,58,66.
We already know variance for this data is 210.88
- ▶ Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.
- ▶ Then the data becomes 27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4
The mean of new dataset is 23.6
- ▶ The sum of squared deviations from mean = 303.68 and the variance = $\frac{303.68}{9} = 33.74$.
- ▶ The standard deviation of the new data set is $\sqrt{33.74} = 5.808$.
We can verify $5.808 = 0.4 \times 14.522$

So, when I multiply with a constant, I know new variance = $c^2 \times$ old variance. So, when I look at the dataset, you can see that the standard deviation is 0.4×14.522 . So, new standard deviation = $c \times$ old standard deviation this is my c and I can see that, that is the c times the old standard deviation which is 14.522.

(Refer Slide Time: 29:33)

Statistics for Data Science -I
└ Numerical summaries
 └ Measures of dispersion

Section summary



- ▶ Measures of dispersion
 - 1. Range ✓
 - 2. Variance: population variance and sample variance.
 - 3. Standard deviation ↗
- ▶ Impact of adding a constant or multiplying with a constant on the measures.



So, what let us look at this in our. So, you can see that, when I look multiply it with a constant, the standard deviation of my original dataset is 14.522. This you can verify is nothing, but my square root of the sample variance. So, square root of the sample variance is 14.522. In the Google sheets, the command is STDEV standard deviation of B2 to B11 gives the standard deviation.

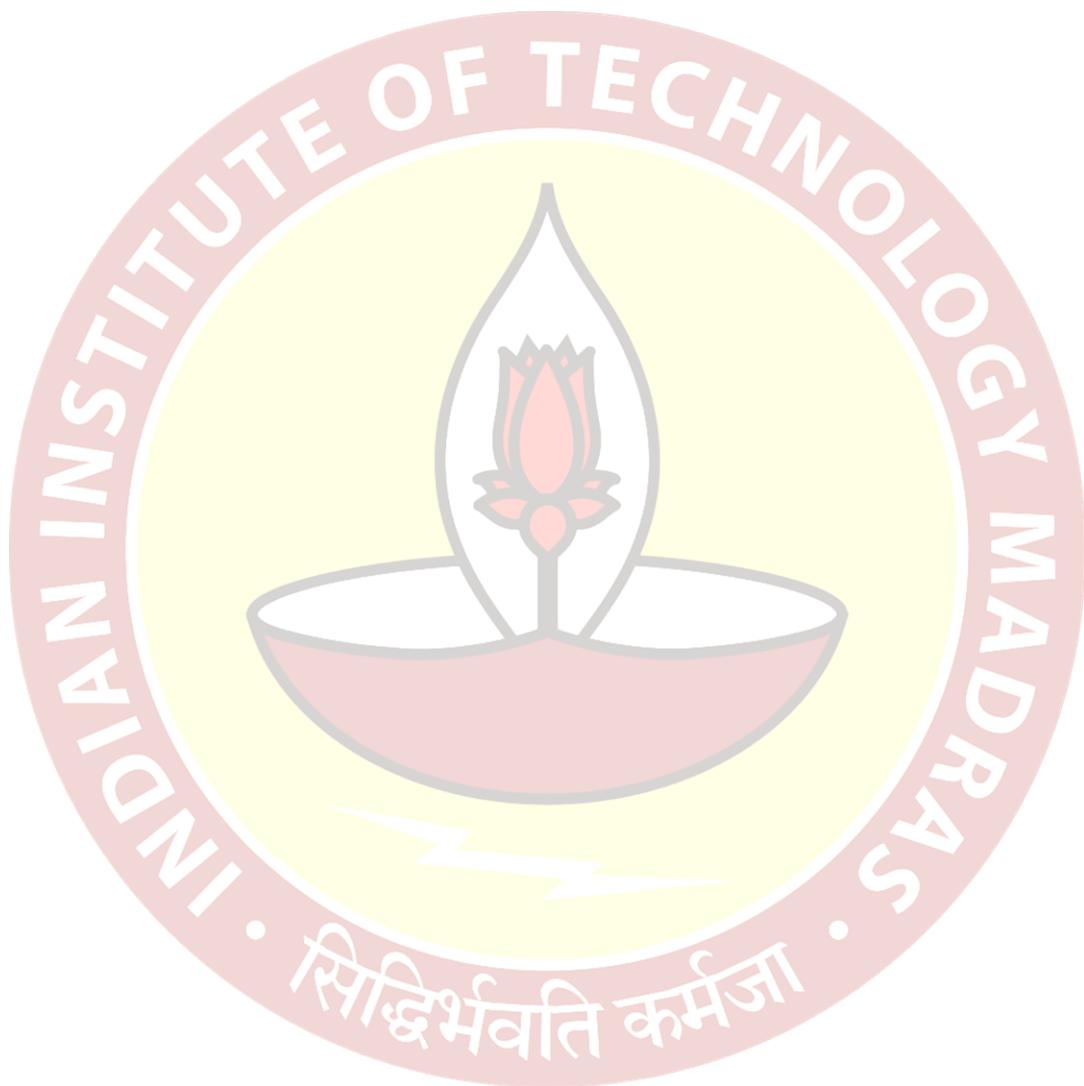
I can see that when I add a constant, the standard deviation remains the same. You can see that the standard deviation for these two, the highlighted columns are the same whereas, when I multiply it with a constant, you can see that the standard deviation is 5.808. You can verify this is 0.4 times my standard deviation is 5.808 and you can see that is equal to my standard deviation of multiplying with the constant.

Hence, when we multiply it with a constant, you can see that the standard deviation, the new standard deviation is the constant times your old standard deviation, okay. So, adding a constant does not change the variability of a set, multiplying with a constant changes the variability of a set by a scalar multiple.

So, what we have seen so far is we started with range, we saw the range is very sensitive to outliers. We defined what was population variance and sample variance. The definitions are very important because just by talking about variance, we need to know whether we are referring to population or sample variance. The numerator captures the sum of square deviations; the denominator is a number of observations if you are

considering population variance and number of observations - 1 if you are considering the sample variance.

Then, we introduce the notion of a standard deviation which has the same units as the original data. We also saw that when you add a constant variability of a dataset does not change whereas, when you multiply with a constant the variability changes.



Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 3.5
Describing Numerical Data – Percentiles, Quartiles, and Interquartile range

(Refer Slide Time: 00:14)

Statistics for Data Science - 1
└ Numerical summaries
└ Percentiles

Percentiles

Percentile: Median

100p 50% 50%

► The sample $100p$ percentile is that data value having the property that at least $100p$ percent of the data are less than or equal to it and at least $100(1-p)$ percent of the data values are greater than or equal to it.

Median

Figure source: Mann, P. S. (2007). Introductory statistics. John Wiley & Sons

The next important thing which we are going to discuss is the notion of what we call a Percentile. These days with a lot of competitive examinations, most of these course and competitive examinations are reported as percentiles. Percentile is different from percentage.

What is a percentile? The sample 100 percent percentile is that data value that has the property that at least $100p$ percent of the data are less than or equal to it and at least $100(1-p)$ percent of the data are greater than or equal to it.

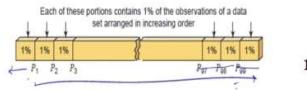
So, now if $p = 1/2$, $100p = 50\%$ of the data, $(1 - p) = 1/2$, $100(1 - p) = 50\%$ of the data and we have already seen a measure which says that 50 percent of the data is less than the value and 50 percent of the data is greater than the value and we call this as the median of a dataset. So, the $100 * 1/2$ or the 50th percentile is the median of the dataset.

(Refer Slide Time: 01:59)



Percentiles

- The sample $100p$ percentile is that data value having the property that at least $100p$ percent of the data are less than or equal to it and at least $100(1 - p)$ percent of the data values are greater than or equal to it.



1



¹Figure source: Mann, P. S. (2007). Introductory statistics. John Wiley & Sons

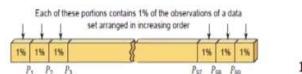
So, to demonstrate it using a figure, you can see that 99th percentile would have $100 * 0.99$ that is 99% of the data is less than it, but 1% is greater than it. Similarly, P_1 says 1% is less than it whereas, 99% is greater than or equal to it. So, the concept of the percentile tells us that value in the dataset below which I have $100 * p$ which are less than or equal and $100 * (1 - p)$ which are greater than or equal to it .

(Refer Slide Time: 02:38)



Percentiles

- The sample $100p$ percentile is that data value having the property that at least $100p$ percent of the data are less than or equal to it and at least $100(1 - p)$ percent of the data values are greater than or equal to it.



1



¹Figure source: Mann, P. S. (2007). Introductory statistics. John Wiley & Sons

Now, if two data values satisfy the same the condition, it is the arithmetic average of these values we have already seen how to compute the median when I have odd or even number of observations.

(Refer Slide Time: 02:55)

Statistics for Data Science -1
└ Numerical summaries
└ Percentiles

Computing Percentile

To find the sample $100p$ percentile of a data set of size n

$\frac{1}{2}$

/1. Arrange the data in increasing order.
2. If np is not an integer, determine the smallest integer greater than np . The data value in that position is the sample $100p$ percentile.

How do I compute a percentile? There are many algorithms to compute a percentile. I now present with a very simple algorithm to computer a percentile which is also very commonly used algorithm.

So, suppose I have a data of size n that is I have n observation, I want to find out what is my sample $100p$ percentile, what I do is I arrange the data in ascending order that is my first step similar to what we did when we computed a median. Once, I arrange my data in ascending order, I find out what $n * p$ is.

So, now let me give you the analogy of what we did with respect to the median. So, I arrange the data in ascending order, this is something which I did for the median also, remember $p = \frac{1}{2}$ so, I check what is $n/2$. If n is even, I know $n/2$ would be an integer. If n is odd, I know $n/2$ is not an integer. So, if np is not an integer, I determine the smallest integer greater than or equal to np that is the value, in that position is the sample $100p$ percentile.

(Refer Slide Time: 04:21)

Statistics for Data Science -1
 └ Numerical summaries
 └ Percentiles

Computing Percentile

$n=5 \quad p=1/2 \quad np = \frac{5}{2} = 2.5 = 3$

To find the sample 100p percentile of a data set of size n

1. Arrange the data in increasing order.
2. If np is not an integer, determine the smallest integer greater than np . The data value in that position is the sample 100p percentile.
3. If np is an integer, then the average of the values in positions np and $np + 1$ is the sample 100p percentile.

$n=6 \quad p=1/2 \quad np = \frac{6}{2} = 3$



If np is an integer, then I look at the average of these values in positions np and $np + 1$. So, what did I say if n/p is not an integer for example, if I look at $n / 2$, suppose $n = 5$, my $n / 2$ is 2.5, the smallest integer greater than this is 3 so, you saw that in a dataset of 5 points, the third value would give me my median. So, that corresponds to my algorithm here.

I repeat, if np is not an integer, let $n = 5$, $p = 1/2$, $n * p = 5 / 2$ which is 2.5 it is not an integer, the smallest integer greater than this is 3. So, the value or the data which is in the 3rd position because I am already arranging my data in ascending order will give me the sample $100 * p$ or the 50th percentile or the median.

But if np is an integer. So, now, let $n = 6$ and $p = 1/2$, my np is an integer which = 3, then what do I do? I look at the average of the value so, the 3rd + the 4th data value that is what we do and I have six data points x_1, x_2, x_3, x_4, x_5 and x_6 which are arranged in ascending order, then you know that $(x_3 + x_4) / 2$ is my median this is how we define. So, I do the same thing if np is an integer, I look at the average of values in positions np and $np + 1$.

(Refer Slide Time: 06:27)

Statistics for Data Science - I
└ Numerical summaries
└ Percentiles

Example



Let $n = 10$

- Arrange data in ascending order 35, 38, 47, 58, 61, 66, 68,
68, 70, 79
 $\leq 9 \leq 10$

p	np	
0.1	1	$(35+38)/2=36.5$
0.25	2.5	47
0.5	5	$(61+66)/2=63.5$
0.75	7.5	68
1	10	79

So, let us look at $n = 10$. I have my data. I again arrange my data in ascending order this is the same data we have been using. So, I have arranged my data, my $n = 10$. So, for $p = 0.1$, my $n * p$ is $10 * 0.1$ which is 1, it is an integer.

So, going back to your algorithm, I need to look at the value which is np and $np + 1$, the 1st value np is 35, $np + 1$ is 38. So, $(35 + 38) / 2$ is 36.5, this is my 100 so, you look at 0.25, 2.5 it is not an integer so, the 3rd value which is 47 that gives me the 25th percentile.

0.5 we have already seen 63.5, again 61 which is my 5th value 1, 2, 3, 4, 5 $(61 + 66) / 2$ that would give me 63.5. 0.75 is again a fraction so, you can see that when I have 0.75, 7.5 so, the 8th value so, this is the 1st, 2nd, 3rd, 4th, 5th, 6th, 7th and 8th, 9, 10, 8th value is 68, 68 is the 75th percentile, the 100th percentile is the maximum which is 79. So, this is how you compute percentiles and I the example shows how to compute percentile for the given dataset.

(Refer Slide Time: 08:14)

Statistics for Data Science -1
└ Numerical summaries
└ Percentiles



Computing percentile using googlesheets-PERCENTILE function

Step 1 Paste the dataset in a column.

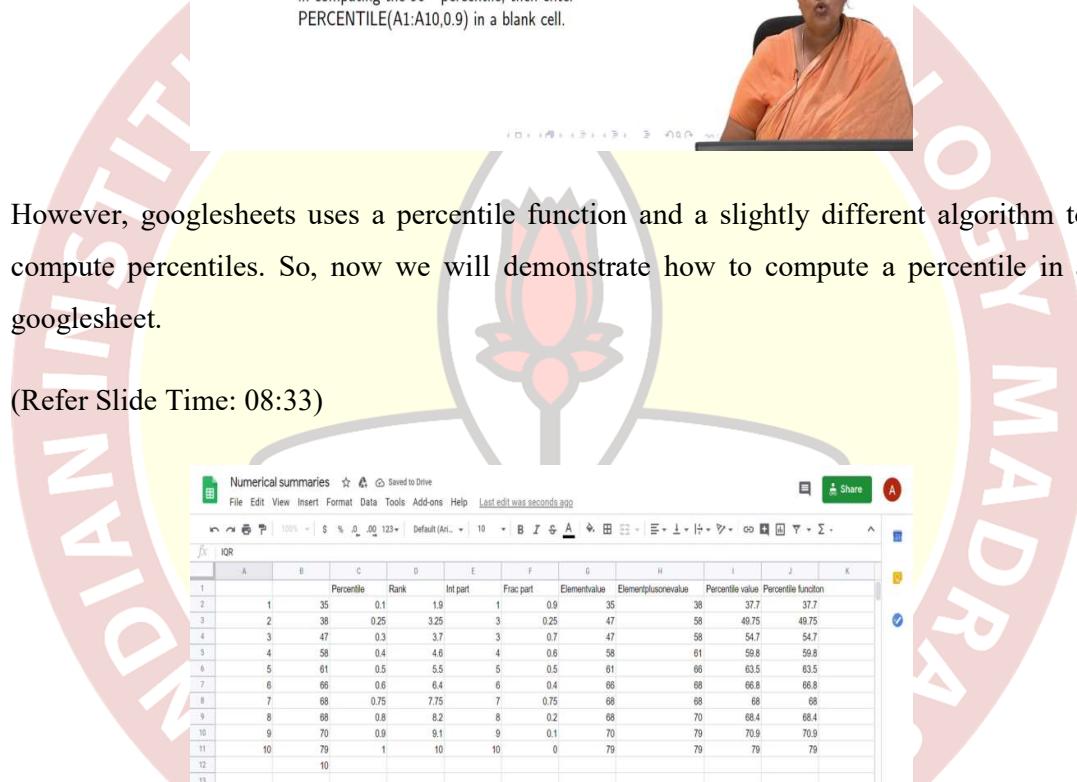
Step 2 In a blank cell enter PERCENTILE(data, percentile), where data indicates the range of data for which percentile needs to be computed, and percentile is the decimal form of the desired percentile.

- For example if the data is in cell A1:A10, and we are interested in computing the 90th percentile, then enter PERCENTILE(A1:A10,0.9) in a blank cell.



However, googlesheets uses a percentile function and a slightly different algorithm to compute percentiles. So, now we will demonstrate how to compute a percentile in a googlesheet.

(Refer Slide Time: 08:33)



IQR											
A	B	C	D	E	F	G	H	I	J	K	
1		Percentile	Rank	Int part	Frac part	Elementvalue	Elementplusonevalue	Percentile value	Percentile function		
2	1	35	0.1	1.9	1	0.9	35	38	37.7	37.7	
3	2	38	0.25	3.25	3	0.25	47	58	49.75	49.75	
4	3	47	0.3	3.7	3	0.7	47	58	54.7	54.7	
5	4	58	0.4	4.6	4	0.6	58	61	59.8	59.8	
6	5	61	0.5	5.5	5	0.5	61	66	63.5	63.5	
7	6	66	0.6	6.4	6	0.4	66	68	68.8	68.8	
8	7	68	0.75	7.75	7	0.75	68	68	68	68	
9	8	68	0.8	8.2	8	0.2	68	70	68.4	68.4	
10	9	70	0.9	9.1	9	0.1	70	79	70.9	70.9	
11	10	79	1	10	10	0	79	79	79	79	
12		10									
13											
14											
15											
16											
17	Q1										
18	Q3										
19	IQR										
20											
21											
22											
23											
24											

So, let us go back to this google sheet. I have the following dataset the same dataset which has been given in ascending order. This is the dataset which I have written the dataset in ascending order.

In a blank cell enter percentile and the data. So, I just go here, I type out what is the percentile. So, if I am looking at this dataset, I look at the percentile, so, B2 to B11 that is my dataset with C2, C2 is 0.1 so, this gives me what is a percentile so, for example, if the data we are interested in 90th percentile, I put a 0.9, if I am interested in the 10th percentile, I put a 0.1. So, you can see that 37.7, the 90th percentile I am putting it, here is 70.9, the 100th percentile is 79. So, you can see this is how we have computed or the google sheet computes it.

Again I repeat, the way google sheet you do is choose the data that is B2 to B11 that is the data which I want to compute the percentile for and C2, C2 is 0.1, 0.25 will give me the 25th percentile this gives me the 10th percentile, 54.7 will give me the 30th percentile, 0.9 gives me 70.9 is the 90th percentile and 79 is the 100th percentile.

Notice that the percentiles need not be part of the dataset. What you will notice immediately is these percentiles which google sheet percentile function gives us is different from the for the same dataset it is different.

Here I got my 10th percentile to be 36.5 whereas, google sheet gives me 37.7. So, the algorithm that google sheet uses, it is not that this is wrong and that is right, but the algorithms use as I mentioned earlier the algorithms used are different.

(Refer Slide Time: 11:04)

STATISTICS FOR DATA SCIENCE - I

- └ Numerical summaries
- └ Percentiles

Computing percentile using googlesheets-algorithm

Step 1 Arrange data in increasing order.

Order	1	2	3	4	5	6	7	8	9	10
$x_{[i]}$	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
Data	35	38	47	58	61	66	68	68	70	79

Let $x_{[i]}$ denote the i^{th} ordered value of the dataset.

Step 2 Find rank using the following formula.

$rank = percentile \times (n - 1) + 1$ where n is total number of observations in the dataset

10 9



So, what is the algorithm a google sheet uses to compute percentile? So, first it arranges the data in ascending order. So, I have the data these are $x_1, 2, x_3$ these are the ranks the data is arranged in ascending order that is what we have done here the data is arranged in ascending order.

The second step is for any observation, x_i is denoting the i th order value, find the rank using the following formula. So, the rank is percentile * $(n - 1) + 1$ where n is the total number of observations.

(Refer Slide Time: 11:54)

Computing percentile using googlesheets-algorithm

Step 1 Arrange data in increasing order.

Order	1	2	3	4	5	6	7	8	9	10
$x_{[i]}$	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
Data	35	38	47	58	61	66	68	68	70	79

Let $x_{[i]}$ denote the i^{th} ordered value of the dataset.

Step 2 Find rank using the following formula.
 $rank = percentile \times (n - 1) + 1$ where n is total number of observations in the dataset

- Example: to compute 25 percentile of a set of $n = 10$ observations, $rank = 0.25 \times (10 - 1) + 1 = 3.25$

Step 3 Split the rank into integer part and fractional part.

- Integer part of 3.25 = 3; fractional part is 0.25.

Step 4 Compute the ordered data value $x_{[i]}$ corresponding to the integer part rank.

A photograph of a woman in an orange sari sitting at a desk, looking towards the camera. There is a small diagram next to her showing a fraction 3 over 0.25.

So, let us look at it n in this case is 10, $10 - 1$ is 9 ok, percentile * $(n - 1)$. So, for example, if I want the 25th percentile, I put $0.25 * 10 - 1$ so, I have $0.25 * 9 + 1$. I get a 3.25, is it clear.

So, you can see that here what I have done is the computed the rank using that same formula. So, what is this formula? It is C2, C2 is your percentile * $(n - 1)$ which is my 9 + 1. So, this is the data. So, percentile into so, I am computing the rank of each of these datasets. So, the rank here is 1.9, 3.25 we have already demonstrated how I got this 3.25.

For each of these ranks, I have an integer part and I have a fractional part. If the rank is 1.9, the integer part is 1, the fractional part is 3.9. For 3.25, the integer part is 3, the fractional part is 0.25 ok.

So, for each one of them, I have an integer part and a fractional part. So, in the 3rd step, I split it into an integer and fractional part and that is what we have done here. For the rank, I have split it into an integer part and the fractional part. For every rank, first we compute the rank, then I split the rank into integer and fractional part.

Once that is done, I look at what is the I compute the ordered data value corresponding to the integer part rank. So, what do I mean by this? For 3.25, the integer part is 3, the fractional part is 0.25.

(Refer Slide Time: 14:05)

Statistics for Data Science - I
└ Numerical summaries
└ Percentiles

Computing percentile using googlesheets-algorithm

Step 1 Arrange data in increasing order.

Order	1	2	3	4	5	6	7	8	9	10
$x_{[i]}$	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
Data	35	38	47	58	61	66	68	68	70	79

Let $x_{[i]}$ denote the i^{th} ordered value of the dataset.

Step 2 Find rank using the following formula.
 $rank = percentile \times (n - 1) + 1$ where n is total number of observations in the dataset

- Example: to compute 25 percentile of a set of $n = 10$ observations, $rank = 0.25 \times (10 - 1) + 1 = 3.25$

Step 3 Split the rank into integer part and fractional part.

- Integer part of 3.25 = 3; fractional part is 0.25.

Step 4 Compute the ordered data value $x_{[i]}$ corresponding to the integer part rank.

- The ordered data value corresponding to integer part rank of 3, $x_{[3]}$ is 47.

3 | 2 25
| 0 50

A photograph of a woman in an orange sari sitting at a desk, looking towards the camera. A small video player interface is visible at the bottom of the slide.

Compute the ordered data value corresponding to the integer part rank. The integer part rank is 3 and you can see that the ordered data corresponding to this integer part rank is 47.

(Refer Slide Time: 14:26)



Computing percentile using googlesheets-algorithm

Step 1 Arrange data in increasing order.

Order	1	2	3	4	5	6	7	8	9	10
$x_{[i]}$	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
Data	35	38	47	58	61	66	68	68	70	79

Let $x_{[i]}$ denote the i^{th} ordered value of the dataset.

Step 2 Find rank using the following formula.

$$\text{rank} = \text{percentile} \times (n - 1) + 1 \text{ where } n \text{ is total number of observations in the dataset}$$

- Example: to compute 25 percentile of a set of $n = 10$ observations, $\text{rank} = 0.25 \times (10 - 1) + 1 = 3.25$

Step 3 Split the rank into integer part and fractional part.

- Integer part of 3.25 = 3; fractional part is 0.25.

Step 4 Compute the ordered data value $x_{[i]}$ corresponding to the integer part rank.

- The ordered data value corresponding to integer part rank of 3, $x_{[3]}$ is 47.

$$3 \frac{0.25}{0.25}$$



Again, for this 25th percentile 3.25, integer part is 3, the fractional part is 0.25. So, what I do is I look at computing or I look at that value which is corresponding to the integer part rank which is 47 that is what we have done here ok.

So, I have this element value. So, 1, 35 the integer value corresponding to 1 is 35, corresponding to 3 is 47, 4 is 58, 5 is 61, 6 is 66, 7 is 68 again here, I have the integer part which is 8, again it is 68, with 9 it is 70 and with 10 it is 79. This is the value which corresponds to the integer. So, 1 you can see its 35, 3 it is 47 and so forth.

(Refer Slide Time: 15:22)



Computing percentile using googlesheets-algorithm-contd

Step 5 The percentile value is given by the formula

$$\text{Percentile} = x_{[i]} + \text{fractional part} \times [x_{[i+1]} - x_{[i]}]$$

- $\text{Percentile} = 47 + 0.25 \times [58 - 47] = 47 + 0.25 \times 11 = 47 + 2.75 = 49.75$

Once that is done, in the 5th step you find out what is the percentile value. The percentile value is the x_i which is $47 + \text{the fractional part}$, remember the part 25th percentile was 3.25, the fractional part was 0.25, the integer part was 3, the value corresponding or the ordered corresponding to x_3 was 47 I take the fractional point which is 0.25 and I look at $x_i + 1$. What is $x_i + 1$? It is 58.

So, I look at $x_i + 1$ which is 58, $- x_i$ which is - 47. So, $47 + 0.25 * [58 - 47]$ is 49.75 and that is what I have here. In the first thing, I have again element value is 35, element value + 1 is 38 which is x_2 , $x_2 - x_1$ is what I have is 2, I multiply that with my fractional part that is what you do here which is 0.9. I add that and I get a percentile function which is 37.7. So, I have $33 * 0.9$ which is 2.7, $35 + 2.7$ is 37.7.

So, you can see that this column I gives us the percentile value computed using this equation whereas, column J gives the percentile value computing the percentile function of the google sheets and we can see that both of them are exactly the same.

(Refer Slide Time: 17:31)

Statistics for Data Science -1
└ Numerical summaries
└ Percentiles

Quartiles	$Q_1 \rightarrow$ First (LOWER) $Q_2 \rightarrow$ Second MEDIAN $Q_3 \rightarrow$ Third (UPPER)
Definition	The sample 25th percentile is called the first quartile. The sample 50th percentile is called the median or the second quartile. The sample 75th percentile is called the third quartile.

A photograph of a woman in an orange sari sitting at a desk, looking down at her work, is visible in the bottom right corner of the slide.

So, why are percentiles mentioned when we talk about the measures of dispersion? As I earlier mentioned, a very important measure is called the quartile. The sample 25th percentile is called the first quartile, the 50th percentile we already know is the median or the second quartile and the sample 75th percentile is called the third quartile.

So, I have Q1, Q2 and Q3 this is referred to as the first quartile or in some books as the lower quartile, this is the third quartile or referred to as the upper quartile, this is the median or the second quartile which is already we have seen is the median ok.

(Refer Slide Time: 18:22)

STATISTICS FOR DATA SCIENCE

Numerical summaries

Percentiles

Quartiles

Definition

The sample 25th percentile is called the first quartile. The sample 50th percentile is called the median or the second quartile. The sample 75th percentile is called the third quartile.

In other words, the quartiles break up a data set into four parts with about 25 percent of the data values being less than the first(lower) quartile, about 25 percent being between the first and second quartiles, about 25 percent being between the second and third(upper) quartiles, and about 25 percent being larger than the third quartile.

MIN Q1 MEDIAN Q3 MAX

LOGY MADE EASY

30 / 36

Now, when I have the quartiles, the quartiles in fact, break up a dataset into four parts. So, if I have this as my dataset, I know already the median breaks up a dataset into two parts, the first quartile.

The first quartile or the lower quartile I have a upper quartile this is Q2, this breaks up so, I have the least, this is the minimum, I have the maximum, I have the first quartile, the second quartile, the third quartile also referred to as a lower quartile and the upper quartile ok. It breaks up the dataset so, I have part 1, part 2, part 3 and part 4.

So, you can see that the quartiles break up an entire dataset into four parts. 25% of the data lie here, 50% lie here, 75% lie here and entire dataset lies between minimum and maximum.

(Refer Slide Time: 19:34)



The Five Number Summary

- ▶ Minimum ✓
- ▶ Q_1 : First Quartile or lower quartile ✓
- ▶ Q_2 : Second Quartile or Median ✓
- ▶ Q_3 : Third Quartile or upper quartile ✓
- ▶ Maximum ✓



So, why are these again an important measure? If you look at it any descriptive statistics most of the summaries are given / what we refer to as a five-number summary. The five-number summary includes the minimum, the first quartile, the median, the third quartile and the maximum as we have given here. So, this five-number summary is a very good way of summarizing a dataset.

(Refer Slide Time: 20:05)



The Interquartile Range (IQR)

$$\text{IQR} = \frac{\text{Range}}{\text{Max} - \text{Min}} = Q_3 - Q_1$$

Definition

The interquartile range, IQR, is the difference between the first and third quartiles; that is,

$$IQR = Q_3 - Q_1$$

- ▶ IQR for the example



The interquartile range is a very important measure of dispersion which you have we have already seen the range is the difference between maximum and minimum. The

interquartile range is the difference between the third quartile and the first quartile and this is referred to as the interquartile range.

(Refer Slide Time: 20:37)

Statistics for Data Science -1
└ Numerical summaries
└ Percentiles

The Interquartile Range (IQR)

Definition
The interquartile range, IQR, is the difference between the first and third quartiles; that is,

$$IQR = Q_3 - Q_1$$

- ▶ IQR for the example
 - ▶ First quartile, $Q_1 = 49.75$
 - ▶ Third quartile, $Q_3 = 68$
 - ▶ $IQR = Q_3 - Q_1 = 18.25$

So, for our example, the first quartile was 49.75, the third quartile was 68, the interquartile range is 18.25. So, the interquartile range is also a measure of dispersion.

(Refer Slide Time: 20:55)

Statistics for Data Science -1
└ Numerical summaries
└ Percentiles

Section summary

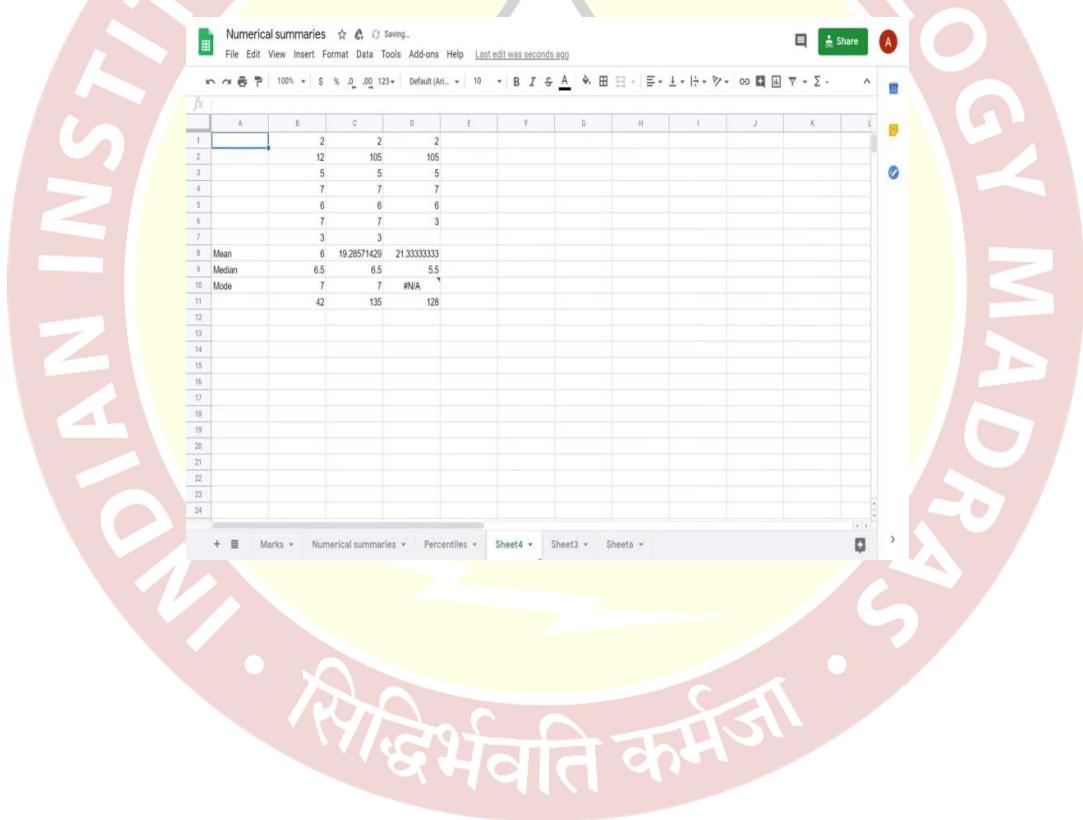
- ▶ Definition of percentiles.
- ▶ How to compute percentiles.
- ▶ Definition of quartile.
- ▶ Five-number summary.
- ▶ Interquartile range as a measure of dispersion.

So, what we have seen so far was how to define percentiles, how to compute percentile, what is the definition of a quartile, the five-number summary which is very very

important in many descriptive statistics and interquartile range which is a measure of dispersion.

The interquartile range so, this if you look at googlesheet, the quartile function with the array and 1 the lower quartile is the first quartile gives me the lower quartile. The quartile function with 3 gives me the upper quartile and the difference between the upper quartile and the lower quartile gives me the interquartile range and we can see that the quartile function if I put 2 gives me the median which you have already seen is 63.5, 3 gives me the upper quartile. So, this is my Q1, this is my Q3 and this is my IQR or my interquartile range.

(Refer Slide Time: 22:08)



(Refer Slide Time: 22:10)

Statistics for Data Science -1

Numerical summaries

Percentiles

Summary

1. Frequency tables
 - 1.1 Frequency table for discrete data. → class
 - 1.2 Frequency table for continuous data. → class intervals
2. Graphical summaries
 - 2.1 Histograms. → class intervals of equal length
 - 2.2 Stem-and-leaf plot. → Mean sensitive outliers
3. Numerical summaries
 - 3.1 Measures of central tendency → Add a const. ↗
3.1.1 Mean, Median, Mode
 - 3.2 Measures of dispersion → Range, sensitive outliers
 - 3.3 Percentiles → Q1 25, SD, Q3 75
 - 3.3.1 Interquartile range as a measure of dispersion.

So, a summary of the this module where we summarize numerical data. What you should be knowing at the end of this module is we first started by looking at frequency tables, we looked at discrete data where each data point was considered as a category or a class, then we looked at group data, we defined what were class intervals, we defined what was lower class limit, upper class limit and we saw how to construct the frequency tables for both discrete and grouped data.

We followed it with histograms. Again, here we assumed class intervals were of equal length and we saw how to construct histograms using google sheets. The stem-and-leaf plot, we also demonstrated how to come up with a stem-and-leaf plot for an example data.

One then, we moved on to numerical summaries. We started with the measure of central tendency although we have looked at mode and median in the earlier categorical data module, but then now we introduce a very important measure called the mean.

What we observed was mean was very sensitive to outliers is something which we saw and then, we saw what would happen to each of these measures of central tendency if you add a constant and if you multiply with the constant. The reason is it is always helps us to know what would happen to the measures whenever we manipulate the data and the way we manipulated the data here was to add a same value or multiply it with the same value.

Then, we moved on to look at measures of variability or dispersion or spread. We started with the measure range, again we saw that the range is extremely sensitive to the outliers we showed this through illustrated it through an example, then we talked about variance.

We also defined the population variance and the sample variance, but we stuck on to the sample variance. At this point of time, we said that all these measures take the same units of your original data whereas, variance takes the units of squared units of the original data hence, we define a measure which is standard deviation which is square root of the variance which takes the units of your original data.

Then, we went on to percentiles. When we discussed about percentile, we introduced what were percentiles, then we introduced important percentiles namely the 25th percentile, the 50th percentile and the 75th percentile.

We call this the first quartile or the lower quartile, the 75th percentile is the 3rd quartile or the upper quartile and the 50th percentile is what we already have seen as the median. The difference between the third and the first quartile is what we refer to as the interquartile range and we see that interquartile range is a measure of dispersion.

So, with this, we come to the end of the discussion on how to come up with numerical summaries for a single variable. What we are going to see next is the association between two variables. So, far we have looked both in the categorical case and the numerical case, measures of graphical summaries and numerical summaries of a single variable.

What would happen when I have more than one variable. So, we look at having two variables, we start with having two categorical variables look at how we, look at the association between two categorical variables, we look at association between 2 numerical variables and then, we looked at association between a categorical and a numerical variable. So, this is what we are going to do next.

Thank you.

Statistics for Data Science – 1
Professor Usha Mohan
Prathyush P(Support team)
Department of Management Studies
Indian Institute of Technology, Madras
Week 3

Tutorial – 01

(Refer Slide Time: 0:14)



Statistics for Data Science - 1
Week 3 Tutorial Questions
Describing numerical data - one variable

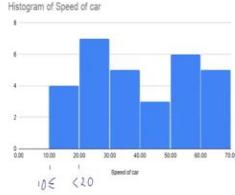
Syllabus covered:

- Visual representation of numerical data and interpret shape of distribution.
- Compute and interpret measures of central tendency: mean, median, mode.
- Compute and interpret measures of dispersion: range, variance, standard deviation.
- Compute and interpret percentiles, Interquartile Range (IQR).
- Compute and interpret five-number summary.
- Use histogram and box-plot to identify outliers in a data set.

Hello statistics students. In this tutorial we want to solve problems based on topics related to week 3. These are the topics, let us look at our first question now.

(Refer Slide Time: 0:28)

↳ A traffic policeman records the speeds of vehicles on a busy road with a 40 kmph speed limit. The histogram represents the results.



- (a) Estimate the average speed of the vehicles.
- (b) Calculate the number of vehicles that were exceeding the speed limit by at least 10 kmph.
- (c) Calculate the number of vehicles having speed greater than or equal to 20 kmph but less than the speed limit.

This is our question number 1. A traffic policeman records the speeds of vehicles on a busy road with a 40 km/h speed limit. The histogram of this data is shown here, this is the histogram of speeds of cars and this is actual speed. So these are the number of vehicles in that class of speeds. So we have 10 to 20, so there are 4 vehicles which went from 10. So we are looking at lesser than, sorry greater than or equal to 10 but lesser than 20. So the left side is what we take as the included boundary. So here let us estimate the average speed of the vehicles.

(Refer Slide Time: 1:33)

↳ A traffic policeman records the speeds of vehicles on a busy road with a **40 kmph speed limit**. The histogram represents the results.

$$\begin{aligned} & 4 \times 15 \\ & + 7 \times 25 \\ & + 5 \times 35 \\ & + 3 \times 45 \\ & + 6 \times 55 \\ & + 5 \times 65 \end{aligned}$$



$$\begin{aligned} & 5 [4 \times 3] + 6 \times 5 + \\ & (5 \times 7) + (3 \times 9) \\ & + (6 \times 11) + \\ & (5 \times 13) \\ = & 5 [12 + 35 + 35 \\ & + 2 + 66 + 65] \\ = & 5 [240] = 1200 \end{aligned}$$

(a) Estimate the average speed of the vehicles. $= \frac{1200}{30} = 40 \text{ kmph}$

(b) Calculate the number of vehicles that were exceeding the speed limit by at least 10 kmph. 11 vehicles

(c) Calculate the number of vehicles having speed greater than or equal to 20 kmph but less than the speed limit. 12 vehicles

For this we are going to use a midpoint of each class, so for this class it is going to be 15, for this one it is going to be 25, for this one it is going to be 35, it is 45, this is 55 and this is 65. And now

how many are there; 4 for the first class, so we are going to do 4×15 plus, this is between 6 and 8, so this must be 7, 7×25 plus, this is 5, we have 5×35 plus this is 3, so 3×45 plus 6, so 6×55 and last one, this is 5. So 5×65 .

$[4 \times 15] + [7 \times 25] + [5 \times 35] + [3 \times 45] + [6 \times 55] + [5 \times 65]$. This is the total calculation we are supposed to be doing.

Taking 5 common out of this stuff, we can write this as

$5([4 \times 3] + [7 \times 5] + [5 \times 7] + [3 \times 9] + [6 \times 11] + [5 \times 13])$. We have slightly more manageable numbers, so we will get $5(12 + 35 + 35 + 27 + 66 + 65)$. So let us sum these up, we will get this is equal to 5 times $35 + 35 = 70$, $70 + 12 = 82$, $82 + 27 = 109$, $109 + 66 = 175$, $175 + 65 = 240$, $5[240] = 1200$. So this is the sum of observations, so all of these put together I will get 1200 divided by, now what is the total number of observations here, that would be these numbers, right? $4 + 7 + 5 + 3 + 6 + 5 = 30$, $4 + 7 = 11$, $11 + 5 = 16$, $16 + 3 = 19$, $19 + 6 = 25$, $25 + 5 = 30$.

So we are getting $\frac{1200}{30}$ and this can get cancelled off to give us 40, so we have 40 km/h is the average, which incidentally is also the speed limit given for these vehicles. So the average is at this speed limit, so the number of them which have crossed the speed limit and that is what the next question is, calculate the number of vehicles that were exceeding the speed limit by at least 10 km/h .

So that would be from 50. So we have interested in the number for this class and this class, which as we can see is 6 and 5, so we get 11, so 11 vehicles. For the last part, we have calculate the number of vehicles having speed greater than or equal to 20 km/h but less than the speed limit. So that would be this and this, this, this, both of these are lesser than 40 but greater than or equal to 20, that gives us 7 + 5 again, so this is then 12 vehicles.

Statistics for Data Science – 1
Professor Usha Mohan
Prathyush P (Support Team)
Department of Management Studies
Indian Institute of Technology, Madras
Week – 03
Tutorial – 02

(Refer Slide Time: 0:14)



This is our second question. This is about a small survey about the data regarding annual fees by some 30 students in 1000s of rupees. So, this will be 20,000 rupees, this will be 40,000 rupees and this would be 1 lakh. Anyway so we are being asked about the correct representation of the survey. There are 4 bar graphs given. We just need to check the counts and which bar graph is representing it properly.

So, let us start with the 20s. 1, 2, 3, 4, 5, 6, 7, there are seven 20s. So, the only options which satisfy this, actually 3 option satisfy it. The option 2, this is definitely wrong because here this is 6 which is wrong and then let us look at 40. 1, 2, 3, 4 and 5. So, which bar graph is giving us 5 for 40? Both a and d, whereas, this is wrong because it is giving 4 which is wrong.

Now, let us look at the 60s, here is 1, 2, 3 and a 4. So, there are four 60s over here. So, certainly a is also wrong because it gives us only 3. So, this is also wrong. So, d must be the answer. Let us

just now verify it. 80; 1, 2, 3, 4, 5 and 6. Are there six 80s? Yes, there are six 80s. We already know that these are all correct and 100s; 1, 2, 3 and 4. So, there are four 100s and yes that is correct. There are four 100s.

And then let us look at 120, 120 there is 1, 2, 3. Three 120s which is also given correct and lastly there is only one 140 which is also represented correctly. So, our correct option is d.

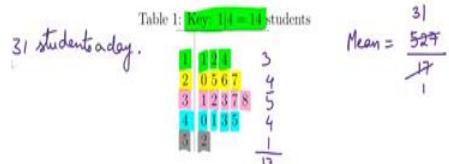


Statistics for Data Science - 1
Professor Usha Mohan
Prathyush P (Support Team)
Department of Management Studies
Indian Institute of Technology, Madras
Week - 03
Tutorial - 03

(Refer Slide Time: 0:14)



Q) A class teacher created the following stem-and-leaf plot showing the attendance of his class over a few days.



Then the average attendance of his class over these days is:

$$\begin{aligned}
 & 11 + 12 + 14 + 20 + 25 + 26 + 27 + 31 + 32 + 33 + 37 + 38 \\
 & + 40 + 41 + 43 + 45 + 52 = 527
 \end{aligned}$$

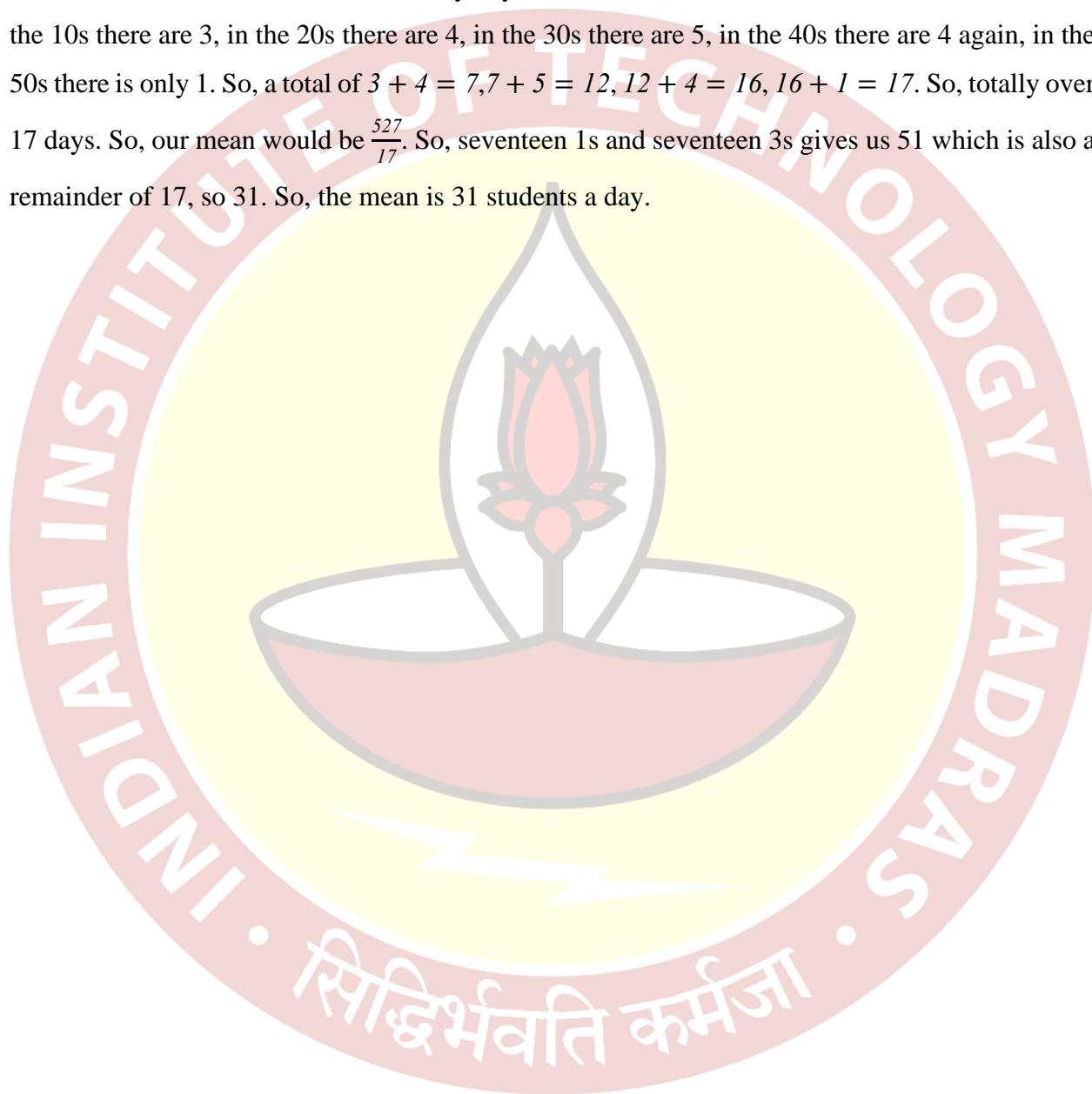
Now third question, there is a stem-and-leaf plot that is shown here. And this is about the attendance of a particular class. There is also a key provider, which shows that 1|4 is essentially 14 students. So, if this is 1 and this is 4, that would imply 14 students on a particular day, they are asking us the average attendance. That would imply we have to do the sum of all the attendances and let us look at the ones first. So, we have 11 and a 12 and a 14, thus we have

$$11 + 12 + 14 + \dots$$

For the 20s we have 25, 26 and 27. So, $20 + 25 + 26 + 27$. Then further on for the 30s we have 31, 32, 33, 37 and 38. So, we have plus $31 + 32 + 33 + 37 + 38$ and the 40s we have 40, 41, 43 and 45. So, that would be plus $40 + 41 + 43 + 45$. And lastly, we have a 52, so plus 52. The total sum of this we will have to calculate and that is going to be

$$11 + 12 = 23, \quad 23 + 14 = 37, \quad 37 + 20 = 57, \quad 57 + 25 = 82, \quad 82 + 26 = 108, \quad 108 + 27 = 135, 135 + 31 = 166, 166 + 32 = 198, 198 + 33 = 231, 231 + 37 = 268, 268 + 38 = 306, 306 + 40 = 346, 346 + 41 = 387, 387 + 43 = 430, 430 + 45 = 475 + 52 = 527.$$

Now, we also have to find out how many days have the attendance has been calculated for. So, in the 10s there are 3, in the 20s there are 4, in the 30s there are 5, in the 40s there are 4 again, in the 50s there is only 1. So, a total of $3 + 4 = 7, 7 + 5 = 12, 12 + 4 = 16, 16 + 1 = 17$. So, totally over 17 days. So, our mean would be $\frac{527}{17}$. So, seventeen 1s and seventeen 3s gives us 51 which is also a remainder of 17, so 31. So, the mean is 31 students a day.



Statistics for Data Science - 1
Professor. Usha Mohan
Prathyush P (Support Team)
Department of Management Studies
Indian Institute of Technology, Madras
Week - 03
Tutorial - 04

(Refer Slide Time: 0:14)

4) The mean of 25 observations is 36. The mean of the first 13 observations is 32 and that of last 13 observations is 39. What is the value of the 13th observation?

22
 33
 26
 23



IIT Madras
ONLINE DEGREE

mean of 32

$t_1, t_2, t_3, \dots, t_{13}, \dots, t_{25}$

$$\frac{\sum_{i=1}^{25} t_i}{25} = 36$$

$$\frac{\sum_{i=1}^{13} t_i}{13} = 32$$

$$\frac{\sum_{i=13}^{25} t_i}{13} = 39$$

$$\sum_{i=m}^n t_i = t_m + t_{m+1} + t_{m+2} + \dots + t_n$$

For our fourth question, we are told that the mean of 25 observations is 36 and the mean of the first 13 observations is 32 and the last 13 observations is 39, which means the thirteenth observation must be included in both of the calculations. Because, so we have observations, let us call them t_1, t_2, t_3 , so on and look at the t_{13} one, it comes right in the middle and then we have till t_{25} .

So, the first 13 would include these and the mean is 32. The last 13 would include these and the mean is 39, so this particular term is there in both calculations. So, what we get from these three different pieces of information is, when you talk about all 25 put together, the mean is 36. So, that

means the total sum $\frac{\sum_{i=1}^{25} t_i}{25} = 36$.

Then we also know that $\frac{\sum_{i=1}^{13} t_i}{13} = 32$ and $\frac{\sum_{i=13}^{25} t_i}{13} = 39$. So, in case you are confused about what this Σ_i going from something to something let us call this m to n t_i means, this is basically the Σ implies a summation so you are adding things. And what are you adding?

You are adding t_i 's, where the i variable goes from m to n. So, that would be t_m plus because you are starting from m and everything on the way till n so, $t_m + t_{m+1} + t_{m+2} + \dots + t_n$, this is what the summation notation indicates.

(Refer Slide Time: 3:04)

$$\begin{aligned}
 & \frac{25}{\sum_{i=1}^{25} t_i = 36} = 32 \\
 & \sum_{i=m}^n t_i = t_m + t_{m+1} + t_{m+2} + \dots + t_n \\
 & \frac{t_1 + t_2 + t_3 + \dots + t_{25}}{25} = 36 \Rightarrow \sum_{i=1}^{25} t_i = 25 \times 36 = 900 \\
 & \sum_{i=1}^{13} t_i = 13 \times 32 = 416. \\
 & \sum_{i=13}^{25} t_i = 13 \times 39 = 507.
 \end{aligned}$$



So, in this particular first case what we are basically saying is $t_1 + t_2 + t_3 + \dots + t_{25}$, that is what it means, where i goes from 1 to 25.

$$\frac{t_1 + t_2 + t_3 + \dots + t_{25}}{25} = 36$$

$$\sum_{i=1}^{25} t_i = 25 \times 36 = 900$$

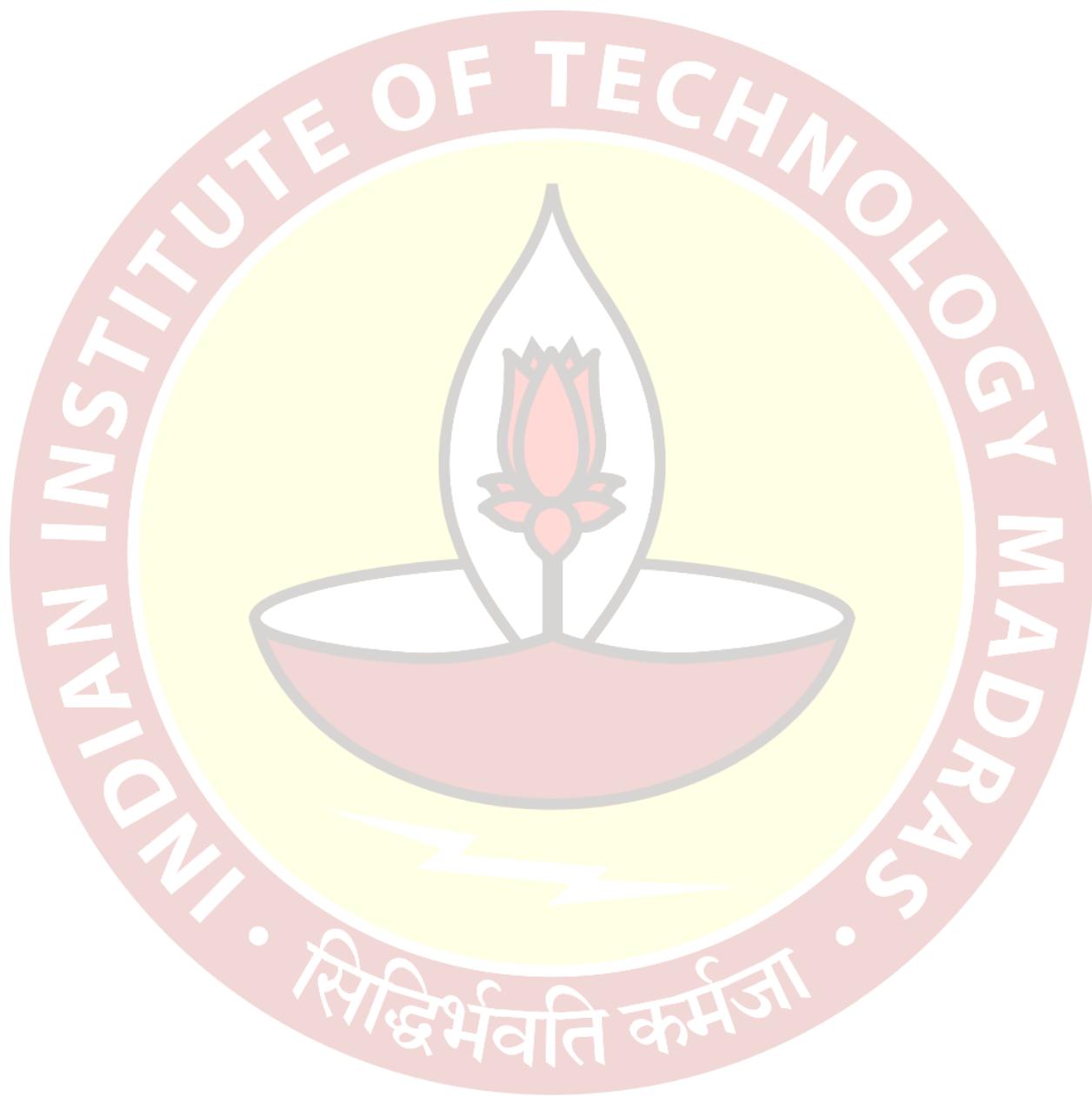
And now from the second piece of information we get that

$\sum_{i=1}^{13} t_i$, that is the sum of the first 13 terms is so what is the mean, the mean is 32, so

$$\sum_{i=1}^{13} t_i = 13 \times 32 = 416.$$

And this portion, the last one where $\sum_{i=13}^{25} t_i = 13$ times because a 13 observation overall into 39, this is the given mean, so this is essentially 507.

$$\sum_{i=13}^{25} t_i = 507$$



(Refer Slide Time: 4:40)

$$\begin{aligned} \rightarrow \sum_{i=1}^{13} t_i &= 13 \times 32 = 416. \\ \rightarrow \sum_{i=13}^{25} t_i &= 13 \times 39 = 507. \\ \sum_{i=1}^{13} t_i + \sum_{i=13}^{25} t_i &= 416 + 507 = 923 \\ \boxed{\sum_{i=1}^{25} t_i} + t_{13} &= 923 \\ \Rightarrow 900 + t_{13} &= 923 \\ \Rightarrow t_{13} &= 923 - 900 = 23 // \end{aligned}$$

So, now what we have is the sum of the first 25 terms is there, sum of the first 13 terms is there and sum of the last 13 terms is there. So, if I added these two, this and this, I will get

$$\sum_{i=1}^{13} t_i + \sum_{i=13}^{25} t_i = 416 + 507 = 923.$$

However, these two put together are basically $\sum_{i=1}^{25} t_i + t_{13}$ which is because our t_{13} is showing up once in both of these summations.

Therefore, when I combine them one t_{13} goes into the total summation and the other extra is lying here, so this is going to give us 923 and this sum, the total sum we know is 900 which means $900 + t_{13} = 923$ and that indicates that $t_{13} = 923 - 900 = 23$.

Statistics for Data Science - 1
Professor Usha Mohan
Prathyush P (Support team)
Department of Management Studies
Indian Institute of Technology, Madras
Week - 03
Tutorial - 05

(Refer Slide Time: 0:14)



5] The following frequency table gives the values obtained in 30 rolls of a die.

Value	Frequency	
1	4	= 4
2	6	= 12
3	7	= 21
4	5	= 25
5	3	= 15
6	5	= 30
		$\sum_{i=1}^{30} (x_i - \bar{x})^2$
		29
		$\bar{x} = \frac{102}{30} = 3.4$
		$\bar{x} = 3.4$
		102

Find

- (a) the sample standard deviation
- (b) the sample variance
- (c) the sample mode. 3

In our fifth question, there is this frequency table given to us and these are values obtained in 30 rolls of a die. So, the value 1 was achieved 4 times, value 2 was achieved 6 times and so on. Now, they want us to find the sample standard deviation, sample variance and sample mode. We should be doing this in the opposite order because the mode is the simplest, so I would like to finish that first, clearly 7 is the maximum frequency and 3 is the value with the maximum frequency.

So, our mode is 3, simple, and then sample standard deviation is simply the square root of the sample variance, so let us calculate the sample variance first and for calculating the sample variance we first have to calculate the mean. So, for calculating the mean, what we need to observe here is the mean is for the values that are being

obtained on the die, which means one has come four times, so that contribution is 1×4 which is 4 but 2 has come 6 times.

So, the contribution from 2 would be 12 and this is going to give us 21. 3 has come 7 times. 4 is going to give us 20. 5 is giving us 15 and lastly 6 is giving us 30. So, this is an individual contributions to the total sum and the total sum would be this, $16 + 21$ is 37, $37 + 20$ is 57, $57 + 15$ is 72, $72 + 30$ is 102. So, the mean is, let us call mean \bar{x} is $102 \div 30$ which gives us this is 3 tens and this is 3×3 and 4. So, 3.4 is the mean, so this is our mean and for calculating variance what we do is we now do $\sum_{i=1}^{30} (x_i - \bar{x})$.

So, for each term we are going to look at the observation, we are going to look at its difference from the deviation from the mean and we are going to square it and divide by 29 that is $30 - 1$, when we are calculating the sample variance and sample standard deviation we use $n - 1$ in the denominator not n .

(Refer Slide Time: 3:24)

$$\text{Find } \bar{x}$$

4	\times	5	$= 20$
5	\times	3	$= 15$
6	\times	5	$= 30$
			$\underline{\underline{102}}$

$$\bar{x} = 3.4$$



(a) the sample standard deviation

(b) the sample variance

(c) the sample mode. 3

$$\frac{\sum_{i=1}^{30} (x_i - \bar{x})^2}{29} = \frac{4(1-3.4)^2 + 6(2-3.4)^2 + 7(3-3.4)^2 + 5(4-3.4)^2 + 3(5-3.4)^2 + 5(6-3.4)^2}{29}$$

So this in order to calculate this quantity, we are essentially going to do $(1 - 3.4)^2$, how many times is this going to come? 4 times. So, $4(1 - 3.4)^2$, plus we have 2 coming up 6 times, so $6(2 - 3.4)^2 + 7(3 - 3.4)^2 + 5(4 - 3.4)^2 + 3(5 - 3.4)^2 + 5(6 - 3.4)^2$. And this whole thing this entire thing is to be divided by 29.

$$\frac{4(1 - 3.4)^2 + 6(2 - 3.4)^2 + 7(3 - 3.4)^2 + 5(4 - 3.4)^2 + 3(5 - 3.4) + 5(6 - 3.4)^2}{29}$$

(Refer Slide Time: 4:35)



$$\begin{aligned}
 &= \frac{(4 \times 5.76) + (6 \times 1.96) + (7 \times 0.16) + (5 \times 0.36) + 3(2.56)}{29} \\
 &\quad + 5(6.76) \\
 &= \frac{23.04 + 11.76 + 1.12 + 1.8 + 7.68 + 33.8}{29} \\
 &= \frac{79.2}{29} = 2.731 \quad \text{Sample Variance} \\
 &\text{Standard Deviation} = \sqrt{2.731} \\
 &\approx 1.6525
 \end{aligned}$$

So, let us calculate that, we get $4[(2.4)^2 = 5.76] + 6[(1.4)^2 = 1.96] + 7[(0.4)^2 = 0.16]$. I will put them in brackets in order to avoid confusion, plus $5[(0.6)^2 = 0.36] + 3[(1.6)^2 = 2.56] + 5[(2.6)^2 = 6.76]$.

Again $\frac{4[5.76] + 6[1.96] + 7[0.16] + 5[0.36] + 3[2.56] + 5[6.76]}{29}$ and simplifying further we get, $4 \times 5.76 = 23.04$ plus $6 \times 1.96 = 11.76$ plus $7 \times 0.16 = 1.12$ plus $5 \times 0.36 = 1.8$ plus $3 \times 2.56 = 7.68$ plus $5 \times 6.76 = 33.8$ divided by the whole thing, divided by 29 which gives us further $23.04 + 11.76 = 34.8$, $34.8 + 1.12 = 35.92$, $35.92 + 1.8 = 37.72$, $37.72 + 7.68 = 45.4$, $45.4 + 33.8 = 79.2$.

$$\frac{79.2}{29} = 2.731$$

This is our sample variance and standard deviation is merely the square root of this, so standard deviation is $\sqrt{2.731}$ which is roughly, this is also roughly I only approximated to 3 decimals anyway this is roughly 1.6525.

(Refer Slide Time: 8:03)

5) The following frequency table gives the values obtained in 30 rolls of a die.

Value	Frequency	
1	4	= 4
2	6	= 12
3	7	= 21
4	5	= 20
5	3	= 15
6	5	= 30
		$\sum 102$

Find

$$\sum_{i=1}^{30} (x_i - \bar{x})^2 = 29$$

$$\bar{x} = \frac{102}{30} = 3.4$$

$$(a) \text{ the sample standard deviation } 1.6525$$

$$(b) \text{ the sample variance } 2.731$$

$$(c) \text{ the sample mode. } 3$$

$$\sum_{i=1}^{30} (x_i - \bar{x})^2 = [4 \times (1 - 3.4)^2 + 6 \times (2 - 3.4)^2 + 7 \times (3 - 3.4)^2 + 5 \times (4 - 3.4)^2 + 3 \times (5 - 3.4)^2 + 5 \times (6 - 3.4)^2]$$

So, our answers would be sample variance is 2.731 and sample standard deviation is 1.6525.

Statistics for Data Science-1
Professor. Usha Mohan
Prathyush P (Support Team)
Department of Management Studies
Indian Institute of Technology, Madras
Week - 03
Tutorial - 06

(Refer Slide Time: 00:16)



Q) Rajan just took his first math test in his college analysis class. His professor says he scored in the 80th percentile for the class. Rajan's professor posts a list of grades, without the names, on the blackboard. There are 12 students in the class and 12 grades on the board. The grades are:
 46, 55, 68, 93, 84, 70, 38, 66, 78, 75, 55, 60. Then the Rajan's grade is

$$\frac{80}{100} \times 12 = \frac{48}{5} = 9.6$$

10th → Rajan scored 84.

In our sixth question, we have Rajan just took his first math test in his college analysis class. His professor says he scored the eightieth percentile and Rajan's professor posts a list of grades without names. The 12 students, these are the grades, these are the marks actually not grades, but anyway. So, what is Rajan's mark? So, there are 12 students and the eightieth percentile would indicate 80 percent of 12, so if I can sell this with 4 I will get 25 with for 4 I will get 3, 55 and 5 16.

So, I get $48/5$, which is equal to 9.6. So, obviously you cannot be the 9.6 person in a class. So, you taken to be the tenth guy, that is in our case we have to order these now. So, in this we will first go look at the smallest value 38 is the smallest followed by 46 which is the second smallest. Then we have a 55 which is the third smallest and then there is 60 the fourth. Sixty sixth the fifth, 68 is the sixth. 70 is the seventh, 75 is the eighth, 78 is the ninth and 84 is the tenth, 85 is eleventh and 93 is of twelfths. So, 84 is our tenth value. So, this implies Rajan scored 84.

Statistics for Data Science – 1
Professor. Usha Mohan
Pratyush P (Support Team)
Department of Management Studies
Indian Institute of Technology, Madras
Week - 03
Tutorial - 07

(Refer Slide Time: 00:14)

Q) Preeti found the following ages (in years) of 8 tigers. Those tigers were randomly selected from the 20 tigers at her local zoo:
 $5, 9, 13, 15, 17, 3, 5, 1$.
Then the value of standard deviation for these 8 tigers' age is

$$\bar{x} = \frac{5+9+13+15+17+3+5+1}{8} = \frac{68}{8} = 8.5$$

$$\sqrt{\frac{\sum_{i=1}^8 (x_i - \bar{x})^2}{7}} = \sqrt{\frac{35^2 + 0.5^2 + 4.5^2 + 6.5^2 + 8.5^2 + 5.5^2 + 3.5^2 + 1^2}{7}}$$

So, I read question. Preeti found the following ages of 8 tigers. So, these are the 8 ages Preeti has found they were selected randomly from 20 tigers at her local zoo, and the value of the standard deviation is being asked. Now this is clearly sample standard deviation, so we need to first find the mean, the mean, \bar{x} would be $5 + 9 + 13 + 15 + 17 + 3 + 5 + 1$, the whole divided by 8, which gives us $5 + 9$ is 14, 14 plus that 13 is 27, 27 + 15 is 42, 42 + 17 is 59, 59 + 3 is 62, 62 + 5 is 67 plus 1 is 68.

So, we are getting $68/8$ which is 8.5. This is our mean. Now, if you look at the deviations of each of these, because we as first going to have to calculate the variance, or we can calculate the standard deviation directly is the square root of the sum of the squares of the deviations. Here, I have used the \bar{x} minus x_i deviation is actually $x_i - \bar{x}$, but since we are squaring, it really does not matter the sign should not matter. So, we can go on like this. Then I goes from 1 to 8 here divided by 7 because it is sample standard deviation.

So, this we will get as square root of the whole thing by 7 in 2. Now, we will get $3.5^2 + 0.5^2 + 4.5^2 + 6.5^2 + 8.5^2 + 5.5^2 + 3.5^2 + 7.5^2$.

(Refer Slide Time: 02:48)

$$\begin{aligned} \sqrt{\frac{\sum_{i=1}^8 (\bar{x} - x_i)^2}{7}} &= \sqrt{\frac{35^2 + 0.5^2 + 4.5^2 + 6.5^2 + 8.5^2 + 5.5^2 + 3.5^2 + 7.5^2}{7}} \\ &= \sqrt{\frac{12.25 + 0.25 + 20.25 + 42.25 + 72.25 + 30.25 + 12.25 + 56.25}{7}} \\ &= \sqrt{\frac{246}{7}} \approx \sqrt{35.143} \approx 5.928 \end{aligned}$$

So, this gives us then the $\sqrt{\frac{12.25 + 0.25 + 20.25 + 42.25 + 72.25 + 30.25 + 12.25 + 56.25}{7}}$ which further gives us square root of this denominators to the 7 the numerator will give us 12.5 + 2.25 is 32.75 plus 42.25 is 75 plus 72.25 is 147.25 plus 30.25 is 177.5 plus 12.25 is 189.75 plus 36.25 is 246, which is roughly the square root of 35.143 which is again roughly 5.928. So, this is the standard deviation of the ages of these tigers, sample standard deviation of course.

(Refer Slide Time: 04:28)

Q) Preeti found the following ages (in years) of 8 tigers. Those tigers were randomly selected from the 20 tigers at her local zoo:

5, 9, 13, 15, 17, 3, 5, 1.

Then the value of standard deviation for these 8 tigers' age is 5.928 years

$$\bar{x} = \frac{5+9+13+15+17+3+5+1}{8} = \frac{68}{8} = 8.5$$

$$\begin{aligned} \sqrt{\frac{\sum_{i=1}^8 (\bar{x}-x_i)^2}{7}} &= \sqrt{\frac{35^2 + 0.25^2 + 4.25^2 + 6.25^2 + 8.25^2 + 5.25^2 + 3.25^2}{7}} \\ &= \sqrt{\frac{12.25 + 0.25 + 20.25 + 42.25 + 72.25 + 30.25 + 12.25 + 56.25}{7}} \end{aligned}$$

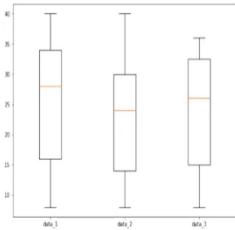
So, our answer is 5.928 years.

Statistics for data Science 1
Professor. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Week 3- Box plot tutorial

(Refer Slide Time: 00:14)



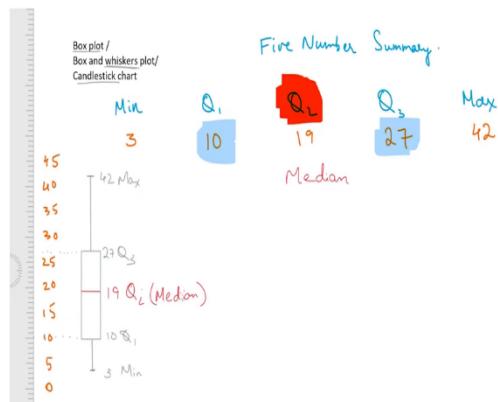
A gardener collected data on rates per kg of tomato sold in different seasons as 12, 16, 22, 8, 36, 34, 40, 32, 28.



Then which of the three box plots shown above correctly represents the gardener's data?

Seventh question, as a gardener who collected data on the rates per kg of tomato, sold in different seasons. So, these are the data points which are 12, 16, 22, 8, 36, 34, 40 and 28. Then the question which of these 3 box plots correctly represents the gardener's data.

(Refer Slide Time: 00:50)



So, for this we have to first understand what a box plot is, so the box plot is called number of things, so it is called a box plot or box and whiskers plot or also candle sticks chart, these are

the various names given to the plot, and it is essentially a chart a plot to represent the five number summary. So, in the five-number summary what do we do we have something which is the minimum of the data, then we have the first quartile, then we have the second quartile and then we have the third quartile and finally we have the maximum of that data.

Let us imagine some data where the minimum is let us say 3. And the maximum is let us say 42 and the Q_1 is occurring at suppose 10, Q_2 is occurring at suppose 19, Q_3 is occurring at suppose 27. So, this is how the data is, we have found the five numbers summary of this data suppose. Now, for the box plot we first establish this markings vertically like this letters in this ruler, let us consider this is 0, this is 5, 10, 15, 20, 25, 30, 35, 40 and 45.

So, this is how the ruler is 0 is here and 45 is here, so what we do in a box plot is we draw a box a rectangular box from the Q_3 to Q_1 so Q_1 to Q_3 we draw a rectangular box which looks like this. So, here we can see that the upper part of the box is Q_3 and the lower part of the box is Q_1 10 to 27. So, this is basically nothing but the interquartile range Q_1 to Q_3 is the interquartile range and we are basically showing a box to represent the interquartile range.

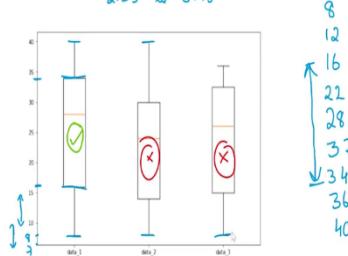
So, this is why we call it the box plot and the whiskers part is that along with this box we also draw a vertical line from the Q_3 to the max here which is 42 and here we mark it off like this. And we do the same thing on the other end from Q_1 to the minimum which is 3, so here this is to 3 so this is our minimum and this is 42 which is our maximum. So, in this way the min is shown the max is shown the Q_1 and Q_3 are shown and what is left is the Q_2 which is also incidentally the median.

So, to indicate the median what we do is we draw this little line that is going through the box. So, this is our 19 which is Q_2 which is also the median because Q_2 is exactly 50 percentile which is a median of the data. And this plot is what is called your box plot or box and whiskers plot or sometimes even the candle sticks plot because it looks like a candle stick. Now, given this introduction, let us go to the question and see how to solve it.

(Refer Slide Time: 05:00)

A gardener collected data on rates per kg of tomato sold in different seasons as 12, 16, 22, 8, 36, 34, 40, 32, 28.

2.25 to 6.75



Then which of the three box plots shown above correctly represents the gardener's data?

So, since it is box plot we need to first rearrange our data as a ascending order we need to arrange the data so we will have 8 first 8 goes first, 12 appears to be next, 16 as after that, 22 comes after that, 28 is here and we have 32, 34, 36 and 40 so overall the 9 observations. So, first of all which box plot has a range of 8 to 40, so this box plot does not seem to be starting from 0 because this appears to be 5 units this length and this is even less than 5 units.

So, maybe this value is probably 7 and then 8 is likely to come about here, so all of these box plots seem to match that the upper limit the other side of the range is 40 for these two. So, either of these two could be our box plot and this one is definitely wrong. Now, let us look at the interquartile range, so we have 9 points we have seen. So, the interquartile range will come from if we did it exactly by as 25 percent, we would get from 2.25 to 6.75 and that means from the third value to the seventh value which is this.

So, 16 into 34 should be shown in our interquartile range. And that is happening for this box plot so this is about at 16 and this is at 34, so this is wrong and this is our correct box plot.

Statistics for Data Science - 1

Prof. Usha Mohan

Department of Management Studies

Indian Institute of Technology, Madras

Lecture No. 4.1

Association between Two Variables - Review of Course

Welcome. This is the week 4 of your online statistics for data science 1 course. In this week, we will understand about association between two variables. So, what are the key things you are going to learn here?

(Refer Slide Time: 00:29)

The screenshot shows a presentation slide with the following content:

Statistics for Data Science - 1

Review

- 1. What is statistics?
 - ▶ Descriptive statistics, inferential statistics.
- 2. Understand how data are collected.
 - ▶ Identify variables and cases (observations) in a data set
- 3. Types of data-
 - ▶ Classify data as categorical or numerical data.
 - ▶ Measurement scales-nominal, ordinal, interval and ratio.
- 4. Describing categorical data
 - ▶ Creating frequency tables, understanding relative frequency
 - ▶ Creating pie charts and bar charts
 - ▶ Descriptive measures of Mode and Median
- 5. Describing numerical data
 - ▶ Creating frequency tables: single valued and grouped data.
 - ▶ Measures of central tendency: Mean, Median, and Mode
 - ▶ Measures of dispersion: Range, Variance, Standard deviation
 - ▶ Percentiles, Quartiles, Interquartile range

On the right side of the slide, there is a small video frame showing a woman in a sari speaking. The slide also features the Indian Institute of Technology Madras logo at the top right.

Before we understand about what we can expect from this week we will just take a quick look at where we stand now. So, we started with understanding what is statistics. Here we actually said that there are two main branches of statistics namely the descriptive statistics and the inferential

statistics part and where our course actually focuses on is in the descriptive statistics and lays the foundation for inferential statistics by introducing probability.

Then we went on to know how data is collected and how data is tabulated and presented. At this point you should know that my columns the way I have constructed a dataset, my columns represent the variables and my rows represent the cases. This is what we should be aware of. Then we went on to classify data mainly when you look at data we can classify it as categorical and numerical. With a numerical you have discrete and continuous.

So, then we saw examples of what we mean by categorical data and what we mean by numerical data. Then we spend some time to understand what were the scales of measurement of variables. Here we introduced 4 prominent scales namely the nominal, ordinal, interval, and ratio scales. The nominal and ordinal scales are used for categorical data. Whereas the interval and ratio scales are used for numerical data.

And during our discussion about scales of measurement we also said what are the arithmetic operations possible for each scale of measurement. Then we went on to describe categorical data. Again we focused on describing categorical data where there is one variable. We started by introducing the notion of frequency table and we introduced the concept of what is relative frequency.

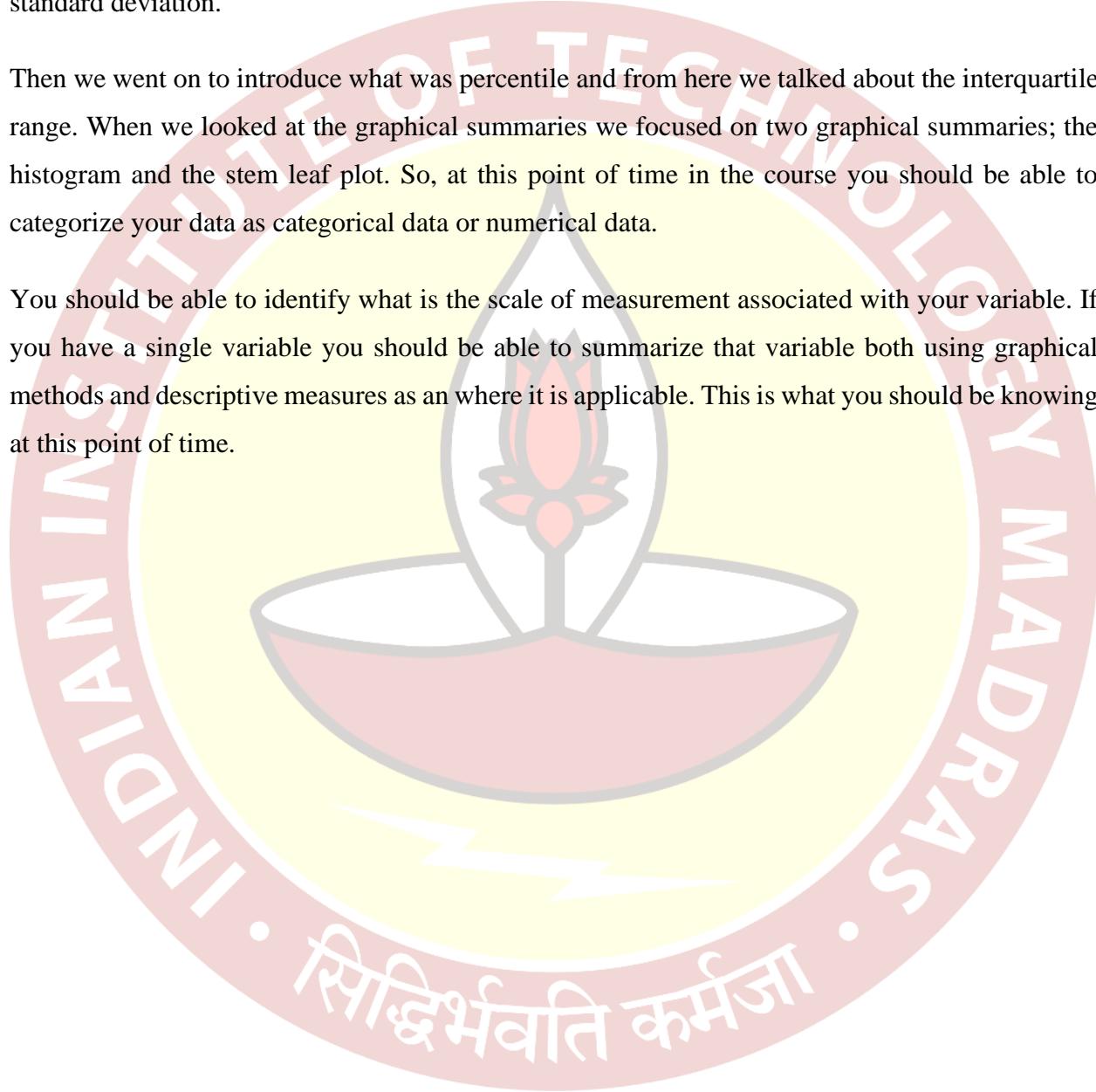
When it came to the graphical measures we talked about creating pie charts and bar charts. Pie charts basically are used when you want to tell the story about what is the share of a particular category in the overall picture whereas bar charts are useful when you want to represent counts. Then we went on to look at descriptive measures of mode and median. While mode can be applied to nominal data also. When you want to talk about a median you want a data to be ordinal or there should be an order in your data.

We then went on to describe numerical data. When we describe numerical we again started by how to create frequency tables. Here we borrowed from our categorical table and we again we talked about a single valued and group data how you create frequency table. Then we introduced the measures of central tendency. These were the numerical measures we introduced a new measure called mean.

When we talk about mean, we want a data to be only interval or ratio. I cannot talk about a mean for a categorical data. We also introduced the notion of median and mode for my numerical data. One important thing which we introduced when we talked about numerical data was measures of dispersion or variation. Here, we started with the range we introduced a measure of variance and standard deviation.

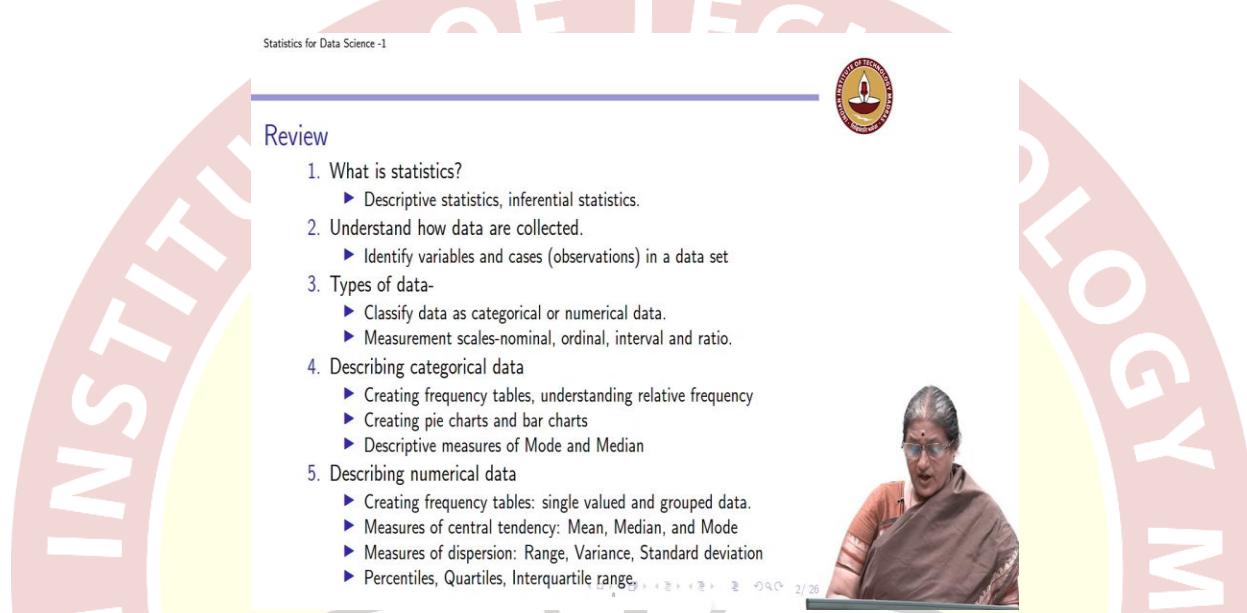
Then we went on to introduce what was percentile and from here we talked about the interquartile range. When we looked at the graphical summaries we focused on two graphical summaries; the histogram and the stem leaf plot. So, at this point of time in the course you should be able to categorize your data as categorical data or numerical data.

You should be able to identify what is the scale of measurement associated with your variable. If you have a single variable you should be able to summarize that variable both using graphical methods and descriptive measures as an where it is applicable. This is what you should be knowing at this point of time.



Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture No. 4.2
Association between Two Categorical Variables - Introduction

(Refer Slide Time: 00:14)



Statistics for Data Science - 1

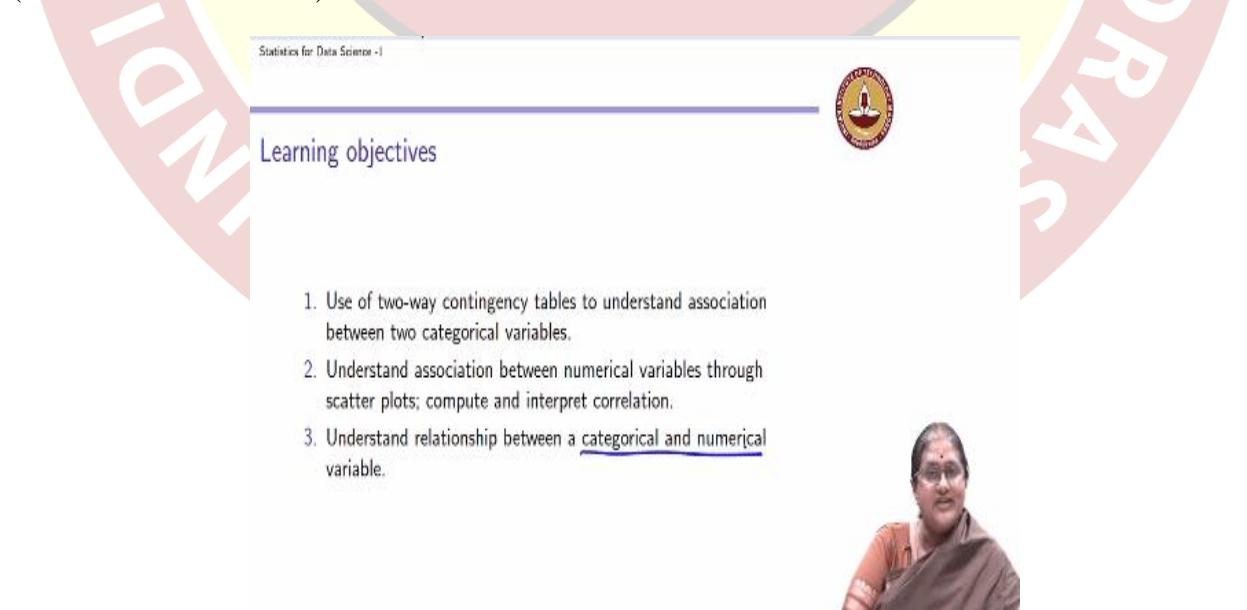
Review

1. What is statistics?
 - ▶ Descriptive statistics, inferential statistics.
2. Understand how data are collected.
 - ▶ Identify variables and cases (observations) in a data set
3. Types of data-
 - ▶ Classify data as categorical or numerical data.
 - ▶ Measurement scales-nominal, ordinal, interval and ratio.
4. Describing categorical data
 - ▶ Creating frequency tables, understanding relative frequency
 - ▶ Creating pie charts and bar charts
 - ▶ Descriptive measures of Mode and Median
5. Describing numerical data
 - ▶ Creating frequency tables: single valued and grouped data.
 - ▶ Measures of central tendency: Mean, Median, and Mode
 - ▶ Measures of dispersion: Range, Variance, Standard deviation
 - ▶ Percentiles, Quartiles, Interquartile range



Now what is the next thing?

(Refer Slide Time: 0:17)



Statistics for Data Science - 1

Learning objectives

1. Use of two-way contingency tables to understand association between two categorical variables.
2. Understand association between numerical variables through scatter plots; compute and interpret correlation.
3. Understand relationship between a categorical and numerical variable.



So far we have focused on understanding only about summarizing a single variable. But most of the time we are interested in understanding whether two variables are associated with each other. When I talk about association I am not mentioning about causality. Association is not always causality. We are not talking about causality here, but you are just asking questions about association between variables.

So, in this module the learning objectives are first we start by understanding association between two categorical variables. Here, we will introduce the notion of contingency tables and how we use contingency tables to understand the association between two categorical variables. Then we move forward to understand how two numerical variables are associated with each other. Here we talk about scatter plots. The nature of this plot and how we measure the association between two numerical variables.

Though the focus of this module is mainly to understand association between two categorical and association between numerical variables, we also spend some time to understand how you will talk about a relationship or an association between a categorical and a numerical variable. So, these are the learning objectives of this week.

(Refer Slide Time: 02:01)

Statistics for Data Science - I
↳ Association between categorical variables
↳ Introduction

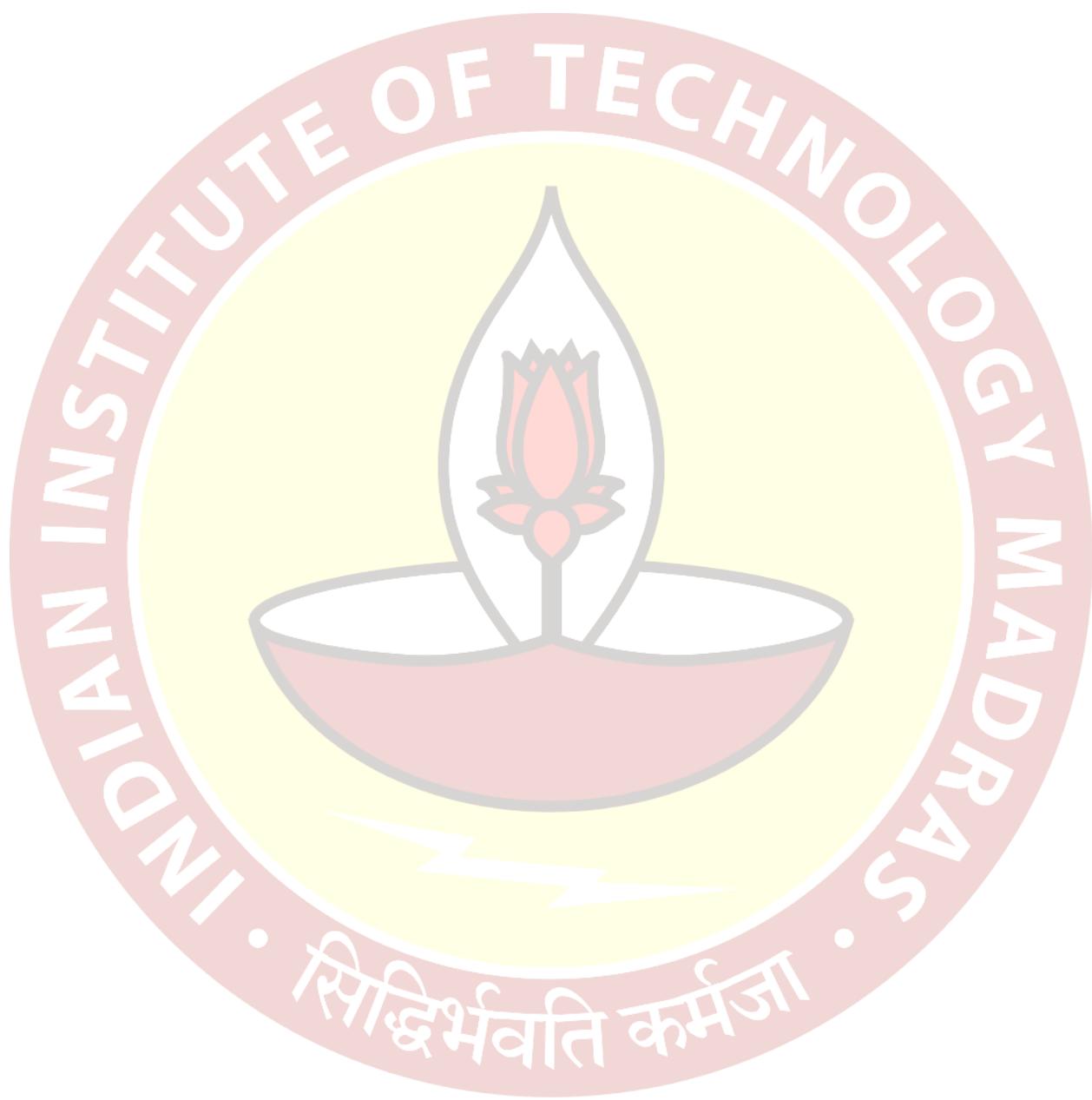
Introduction

▶ To understand the association between two categorical variables.
▶ Learn how to construct two-way contingency table.
▶ Learn concept of relative row/column frequencies and how to use them to determine whether there is an association between the categorical variables.

A portrait of a woman in a sari is visible on the right side of the slide.

So, we start with the association between categorical variables. So, what is the main objective of this section? So, here we are going to understand about how to construct what we refer to as a two

way contingency table. We will introduce the concept of relative frequencies and use how you can use this concept of relative frequency to determine whether there is an association between two categorical variables or not.



(Refer Slide Time: 02:37)



Example 1: Gender versus use of smartphone

- ▶ A market research firm is interested in finding out whether ownership of a smartphone is associated with gender of a student. In other words, they want to find out whether more females own a smartphone while compared to males, or whether owning a smartphone is independent of gender.
- ▶ To answer this question, a group of 100 college going children were surveyed about whether they owned a smart phone or not.
- ▶ The categorical variables in this example are
 - ▶ Gender: Male, Female (2 categories)- Nominal variable
 - ▶ Own a smartphone: Yes, No (2 categories)- Nominal variable



So, let us start with an example. Now if you look at this example I have a market research firm which is interested in finding out whether ownership of a smart phone is associated with gender of a student. In other words, the company is interested in knowing whether more females own a smart phone while compared to males or whether owning a smart phone is independent of the gender of a person. This is the main question.

So, how do we go about answering this question? Immediately you see that when I talk about this I have actually discussed two variables. The first variable is gender and the second variable is ownership of a smart phone. So, how have we captured this gender is again a categorical variable. It has two categories. I am assuming male and female. Ownership the way I have captured the ownership of a smart phone is again through a categorical variable. Here, it is a binary if you own a phone you say yes, if you do not own a phone you say no.

(Refer Slide Time: 04:03)

A screenshot of a Google Sheets document titled "Association between categorical variables". The sheet contains a 2x2 contingency table with gender (M/F) in columns and phone ownership (No/Yes) in rows. The data shows 24 males own phones and 76 do not, while 56 females own phones and 44 do not. Row and column totals are provided, along with relative frequency tables for each gender across both phone ownership categories.

		Gender		Phone Ownership			
		M	F	No	Yes	Total	
1	2	M		No	Yes	10	34
3	3	M		Yes		14	42
4	4	M		Yes		56	
				Grand Total		24	76
						100	
				Row relative frequencies			
				Gender	No	Yes	Grand Total
				F	22.73%	77.27%	46
				M	25.00%	75.00%	56
				Grand Total		24.00%	76.00%
						100	
				Column relative frequencies			
				Gender	No	Yes	Grand Total
				F	41.67%	44.74%	44.00%
				M	58.33%	55.26%	56.00%
				Grand Total		24	76
						100	

So, let us look at the data. So, if you look at the data you can see that this is a data. What is a data? I have the data which has been collected where the variables which I am talking about are basically you can see gender and whether they own a smart phone or not, that is the data that is collected and this data is actually collected for 100 university students. So, what is the data?

The data that is collected is a group of 100 college going children were surveyed about whether they own a smart phone or not. So, the data collected is for each student what was captured, gender and whether they own a phone or not. So yes, no, and gender was captured. So, for example, person 1, I ask a gender it could be a male. This person if they owned a cellphone it would have been yes. Person 2, male does not own a phone.

Person 3 could have been a female does not own a phone. Person 4 could have been a female owns a phone, person n, and this is the way for 100 students we collect this data. So, the two variables here are gender. The second variable here is whether they have a smart phone or not. Yes, if they have a phone no, if they do not have a phone. So, this is the data we have collected. The total number of observations are 100 and this is how we have collected the data.

So, what are the categorical variables in this example? The first categorical variable is gender. They are two categories and we know that gender is a nominal variable. Then next categorical variable is whether you own a smart phone or no. The values this variable take are yes and no. The

values gender take are male and female. Again I have two categories, again it is a nominal variable because there is no order in this variable.

Hence, you can see that I need to know that what is the kind of my variable and what is the scale of measurement. Here I have both the categorical variables with nominal scales of measurement.

(Refer Slide Time: 6:53)

Statistics for Data Science - I
L: Association between categorical variables
Contingency tables

Example 1: Gender versus use of smartphone-summarize data

We have the following summary statistics

1. There are 44 female and 56 male students
2. 76 students owned a smartphone, 24 did not own.
3. 34 female students owned a smartphone, 42 male students owned a smartphone.

The data given in the example can be organized using a two-way table, referred to as a contingency table.

		Own	No Own	Row Total
Gender	Female	10	34	44 ✓
	Male	14	42	56 ✓
Total	24 ✓	76	100	

A hand-drawn 2x2 grid is shown to the left of the table, with 'Gender' at the top and 'Own' at the right. The grid has four boxes: 'Female Own' (top-left), 'Female No Own' (top-right), 'Male Own' (bottom-left), and 'Male No Own' (bottom-right). Handwritten numbers 44, 56, 76, and 24 are placed in these boxes respectively. A blue bracket groups the first three columns under the heading 'Own'.

A woman in a brown sari is visible on the right side of the slide, likely the speaker.

So, now let us look at the data once you have the data, this is the data which I am talking about. So, here you have 100 students and from each of these observations I copy their gender and whether they own a smart phone or not. So, what is the summary statistics I have from my data? The summary statistics I have from my data are I have 44 female students and 56 male students. Remember we are asking 100 students the question a for every student we record what is their gender and we record whether they own a smart phone or not. This is our survey.

Further, 76 students owned a smart phone and 24 did not own. So, if you look at it what we discussed in the previous weeks was, first if I look at gender as a categorical variable how to summarize this gender. I know 44 female and 56 male we saw that we could summarize it using a bar chart. Again owning a smart phone or not. Ownership yes no, I have 76 yes and 24 no. Again there are two values this ownership takes, yes and no and both of them here $44 + 56$ add up to a 100. $76 + 24$ add up to a 100.

Now I want to know the association between these two variables. So, I have another data which is useful to me which says that 34 female students owned a smart phone and 42 owned a smart phone. So, this is the data which is given to us. So, the first question we ask is how do I summarize this data. So, given this data which is in the form of this table which I have here, the question is how do I summarize this data? I have two variables, the first variable I can write it as a gender.

Now there are two values this gender takes I write it as a female, I write it as a male. So this is my variable. The other variable is ownership. Now this is a yes or this is a no this is a yes and I have grand totals. So, if I have the grand row total, I have it here, I have a column total here. So, you can see that there were 44 female that is what is given here, 56 male that is what is given here.

Now when I come to ownership 24 did not own, so my no is 24, 76 owned. So, we can see that this total add up to 100 student. This is the first thing. I am not looking at filling in the inside table, but these are the 44 females, 56 males in my dataset, 24 not owning, 76 owning. Now further the so this is what I have tabulated. So, this is 34 because 34 female owned a smart phone. Similarly, 42 male owned a smart phone.

Now how many did not own that is easy. 10 which is $44 - 34$ and here I have 14 which is $56 - 42$. We can see $10 + 14 = 24$, $34 + 42 = 76$ and we can also check that $10 + 34 = 44$ and $14 + 42 = 56$. So, summarizing this data or the data given here is referred to as a two way table more popularly referred to as a contingency table. How do we construct a contingency table?

To construct a contingency table we look at the first variable. The levels of the first variable in this the first variable is gender and it has two values female and male. So, suppose this variable takes 3 values. So the level 1, level 2, level 3 or level m of the first variable goes into my rows. I look at my second variable. Suppose there are n values of the second variable. I have level 1, level 2, level n, I will have n columns.

And what goes into the i, jth column here is the number of observations of the ith variable and the jth variable together. For example 34 is a number of female students who own a phone. So, this is how we construct what is a contingency table.

(Refer Slide Time: 12:30)

Statistics for Data Science - I
↳ Association between categorical variables
↳ Contingency tables

(Male Fem) N=114

Example 1: Gender versus use of smartphone-summarize data

► We have the following summary statistics Nominal

1. There are 44 female and 56 male students
2. 76 students owned a smartphone, 24 did not own.
3. 34 female students owned a smartphone, 42 male students owned a smartphone.

► The data given in the example can be organized using a two-way table, referred to as a contingency table.

Gender	Own a smartphone		Row total
	No	Yes	
Female	10	34	44
Male	14	42	56
Column total	24	76	100



And you can see that this is how we have summarize the contingency table. Now in this example, both gender and ownership of smart phone both of them where what we referred to as a nominal variable. There was no order in this variable. So, if I had constructed my contingency table by looking at the following, that is male, female, no, yes it would not have made a difference because the information given is the same.

The order did not matter whether I had female and male or yes or no it would not have made a difference. So, the order in which you are stating your variables in the contingency table will not matter when both my variables are nominal in nature.

(Refer Slide Time: 13:39)



Contingency table using google sheets

- Step 1 Choose the columns of the variables for which you seek an association.
- Step 2 Go to Data-click on Pivot table option
- Step 3 Click on create option in the pivot table- it will open the pivot table editor:
 - 3.1 Under the Rows tab, click on the first categorical variable.
 - 3.2 Under the columns tab, click on the second categorical variable.
 - 3.3 Under the values tab, click on either of the variables and then click on the COUNTA tab under "summarize by" tab.



Now we will discuss how to create these contingency table using Google sheet. Go to the data, so this is the data which I have here. You can see that this data has 100 observations on both the variables. The variable gender is in my B column, the variable smart phone is in my C column. So, how do I create the contingency table? I choose my or I highlight my data, how many observations do we have?

We have 100 observations, I highlight my data so I would go and choose the columns. the variables I seek association now are gender and ownership. I choose it then I go to the data tab and click on what is called pivot table. That is step 2, go to data and click on the pivot table option. Now in the pivot table, create pivot table. It opens the pivot table editor in the existing sheet.

I am going to so I go to data I click on pivot table in the existing sheet. I am just going to give a location to my pivot table. The location I give it here and now I go to Create. Under the rows tab add the first categorical variable which is gender. Under the columns tab, click on the second categorical variable which is ownership of a smart phone that is 3.2 step. Under the values tab so I go back here, under the values tab I click on either of the variables.

So, I clicked on gender here and I am asking it to summarize by count A and you can see that I have what I got here is my precisely this is the categorical or the contingency table which we just did a few minutes before. So, you can see that there are 34 females who own a phone, 42 males

who own a phone, 10 females do not own a phone, 14 males do not own a phone, 42 females and 56 males in my dataset and 24 do not own a phone whereas 76 own a phone.

This is what I get from my dataset. So, this is the pivot table in your Google sheet which gives you a contingency table in Google sheet. So, if you saw in the earlier example we had two nominal variables. Now what would happen if I have a ordinal variable?



(Refer Slide Time: 17:02)



Example 2: Income versus use of smartphone

- ▶ A market research firm is interested in finding out whether ownership of a smartphone is associated with income of an individual. In other words, they want to find out whether income is associated with ownership of a smartphone.
- ▶ To answer this question, a group of 100 randomly picked individuals were surveyed about whether they owned a smart phone or not.
- ▶ The categorical variables in this example are
 - ▶ Income: Low, Medium, High (3 categories) - Ordinal variable
 - ▶ Own a smartphone: Yes/No (2 categories) - Nominal variable

LOW
MED
HIGH



Now, let us look at another example here. Now in the earlier thing I saw whether gender was associated with ownership of a phone, I summarize that using a contingency table. Now I am going to look at whether income actually is associated with ownership of a phone. So, again we have the same market research phone which is interested in finding out about whether ownership of a phone again this ownership of a smart phone is my variable here is associated with income of an individual.

Now the income variable is again how do we record this income variable. In this example, we have a market research firm which is interested in finding out whether ownership ownership of a smart phone is associated with income. So, what are the two variables here? The first variable is ownership. Again this was the variable we considered in our earlier example, but now instead of gender I am considering income. Now how is this income recorded?

How is this income recorded, the income is recorded as high, medium or low. So, we have categorized this income into 3 categories and what are the values of this income variable. It is a high, medium and low. So, this is a categorical variable where I am not actually calculating or I have not recorded the actual income, but I have actually categorized these 100 people again into whether they come from a high income group or a medium income group or a low income group.

And for each of these person we are asking whether their income is a high income and whether they own a smart phone or not. This is how I have recorded my data. So, here if you look at this case now what are my variable? Again the categorical variable are income which is low, medium and high and the second categorical variable whether you own a smart phone or not.

This variable whether you own it or not has two categories, the yes category and the no category. It is a nominal variable whereas the income which has three categories, the low, medium and high is an ordinal variable because there is an order in low, medium, and high because low income is lesser than medium income which is lesser than high income. So, recall when you are summarizing two nominal variables.

We said the order in which they appear in the table is not of any relevance. However, when you have an ordinal variable it is good to maintain the order. What do we mean by this?

(Refer Slide Time: 20:25)

A screenshot of a Google Sheets document titled "Association between categorical variables". The sheet displays two contingency tables. The first table shows the relationship between Income (Low, Medium, High) and Own a smartphone (No, Yes). The second table shows the same relationship but with the rows and columns swapped. Both tables include Grand Total rows and columns.

		F	G	H	I	J	K	L
		COUNTA of Inco	CountA Own a smartphone					
		Income	No	Yes	Grand Total			
1								
2								
3		High	2	18	20			
4		Low	9	5	14			
5		Medium	27	39	66			
6		Grand Total	38	62	100			
7								
8								
9								
10								
11								
12								

		F	G	H	I	J	K	L
		COUNTA of Inco	CountA Own a smartphone					
		Income (Coded)	No	Yes	Grand Total			
10								
11								
12								

Suppose I again continue with the same way. I choose income and own a smart phone. I choose these 100 observations, I go to my data, I click on pivot table in my existing sheet I am going to create my pivot table here. Again under rows I add income, under columns I add own a smart phone and under values I am just going to add on a income, you can see that this is my contingency table which I have here, but what you notice in this contingency table is first I have a high which is highlighted here.

So, let us go and look at only the contingency table. So, you can see in the contingency table, if you are looking at the order of the income, you have a high income, a low income and a medium income. Whereas the actual order is either high, medium, low or low medium, high. That is the order in which the variable appears. So, you do not want to see a jumbled order of a variable where there is an order of the variable. One way to overcome this is to have a order, have a high, medium and low variable. I have just coded this variable as 1, 2, 3 where 1 represents a high income, 2 represents a medium income, 3 represents a low income.

So, now if I am looking at a contingency table between these two variables again I choose the 100 observations which I need. I choose the 100 observations. I again go to data, pivot table in the existing sheet I am going to go and create a data a pivot table here. I will just click on this. I am going to create it. Rows I again add the income, columns I add whether they add have a smart phone, values I am going to do the count again.

So, now you can see and you can actually compare these two tables, the first table my high, low and medium did not have an order whereas in the second table, 1 represents a high income group, 2 represents a medium income group, and 3 represents a low income group. So, you can see that the order is preserved in the contingency table. So, whenever you have an ordinal variable it is recommended that the order is preserved in your contingency table.

(Refer Slide Time: 23:58)

Statistics for Data Science - I
↳ Association between categorical variables
↳ Contingency tables

Example 2: Contingency table

- ▶ We have the following summary statistics
 1. There are 20 High income, 66 medium income, and 14 low income participants.
 2. 62 participants owned a smartphone, 38 did not own.
 3. 18 High income participants owned a smartphone, 39 Medium income participants owned a smartphone, and 5 Low income participants owned a smartphone.
- ▶ The contingency table corresponding to the data is given below.

Income level	Own a smartphone		Row total
	No	Yes	
High	12	18	20
Medium	27	39	66
Low	9	5	14
Column total	38	62	100



So, what finally how does my contingency table look for this. I have this data and using this data you can see that the corresponding table, corresponding to this data is I have a high, medium, low the order is preserved in the income whether they own a smart phone or not is recorded. I have 20 people who are high income, 66 who are medium, 14 who are low income. Out of these 162 own a phone, 38 do not own a phone.

Among the high income 20 people, 18 own a phone, 2 do not own a phone. Among the 14, low income group 9 do not have a phone, 5 have a phone. Among the 66, I have 27 who do not own a phone and 39 who own a phone.

(Refer Slide Time: 25:03)

Statistics for Data Science - I
↳ Association between categorical variables
↳ Contingency tables

Section summary

Var 2

Var 1

- ▶ Organize bivariate categorical data into a two-way table - contingency table.
- ▶ If data is ordinal, maintain order of the variable in the table

So, at the end of this subsection you should know how to organize bivariate categorical data into a two way table. Record the first variable and its level in one. These could be the rows. Record the second variable here, this is variable 1, variable 2, and a cell here, the ith level and the jth level tell how many of variable 1 and variable jth level of variable 2, variable 1 ith level and variable 2 jth level are there in this particular cell and this is what is referred to as a contingency table. A word of caution, if the data is ordinal maintain the order of the variable in the table.

Statistics for Data Science -1
Professor. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture No. 4.3
Association between Two Categorical Variables - Relative Frequencies

(Refer Slide Time: 00:14)



The slide title is "Statistics for Data Science -1" with a subtitle "Association between categorical variables" and a sub-subtitle "Relative Frequencies". Below the title is the text "Row relative frequencies". To the right is the Indian Institute of Technology Madras logo.

	No	Yes	Row Total
Female	10/44	34/44	44
Male	14/56	42/56	56
Column Total	24	76	100
	24/100	76/100	

So, we have seen how to construct a contingency table or in other words how to summarize the bivariate data where both my variables are categorical in nature as a contingency table. Now we introduce a very important concept, the concept of a relative frequency. We introduce these concepts when we looked at single categorical variable. Recall relative frequency was nothing, but your frequency by total number of observations.

Relative frequency is what we refer to when we talked about categorical variables as frequency into total number of observations. We have already introduced a notion of a relative frequency, but what do we have in a contingency table. In a contingency table for example I had female, I had male. I had 44 females and I had 56 males, I had 76 people who owned a cellphone.

I have 24 people who did not own a cellphone this was what was my data. So, we can go back to your data here. We are talking about this data here. So, I have 70 in this dataset you can see that I

had 44, 56 then I have 42 men and 34 women owned cellphone. So, 10 women and I have again 14 men did not own a cellphone. So, this is we call this is the contingency table I have.

Now this is my row total what is my row total 44 females and 56 men 24 and my 76 represent or my column totals you can see that the total of the row totals and the total of the column totals are equal and they add up to 100.

(Refer Slide Time: 2:58)

Statistics for Data Science - I
↳ Association between categorical variables
↳ Relative frequencies

Row relative frequencies

► What proportion of total participants own a smart phone?
► What proportion of female participants own a smart phone?

Gender	Own a smartphone		Row total
	No	Yes	
Female	$10/44$	$34/44$	44
Male	$14/56$	$42/56$	56
Column total	$24/100$	$76/100$	100

Row relative frequency: Divide each cell frequency in a row by its row total.

So, now suppose if I am interested in asking a question what is the proportion of total participants who own a phone? Now that is simple total participants are 100 of which 76 people own a phone and 24 people do not own a phone. So, 76 out of 100 people actually own a phone in other words this is the proportion of my total participants who own a phone. Similarly, 24 out of 100 people do not own a phone. So this answer to this question is easily given that the proportion of total participants who own a phone is 76 percentage.

Now let me modify this question a bit and ask that what is the proportion of female participants who own a phone. Now how do we answer this question again if you go back here you can see that there are totally 44 female participants of which 34 participants own a phone. So, the proportion of female participants who own a phone is 34 divided by 44 of the total number of female participants how many female participants we have totally 44 of them I am asking what is the proportion who own a phone 34 by 44.

Similarly, I will have 10 by 44 is the proportion of female participants who do not own a phone. Likewise, 14 by 56 is a proportion of male participants who do not own a phone and 42 by 56 is a proportion of male participants who own a phone. Now what is this 10 by 44? 34 by 44, 14 by 56 and 42 by 56? These are what we refer to as the row relative frequency. What is a row relative frequency? I divide each cell frequency by its row total so I divide this by 44, this by 44, this by 56 and this by 56. Of course for the column this total also 24 by 100 and 76 by 100. So, I know 76% of my total participants owned a phone and 24% of my total participants did not own a phone.

(Refer Slide Time: 05:57)

Statistics for Data Science -I
L- Association between categorical variables
↳ Relative frequencies

Example 1: Row relative frequency

Gender	Own a smartphone		Row total
	No	Yes	
Female	10/44	34/44	44
Male	14/56	42/56	56
Column total	24/100	76/100	100

Gender	Own a smartphone		Row total
	No	Yes	
Female	22.73%	77.27%✓	44
Male	25.00%	75.00%✓	56
Column total	24.00%	76.00%✓	100



And once I do this I see that these are in percentages. 76.27% of the total female participants own a phone that is what this number gives me. 75% of total male participants own a phone, 76% of the total participants own a phone. So, what you have in these cells are what are referred to as the row relative frequencies.

(Refer Slide Time: 06:37)



Example 2: Row relative frequency

Income level	Own a smartphone		Row total
	No	Yes	
High	2/20	18/20	20
Medium	27/66	39/66	66
Low	9/14	5/14	14
Column total	38/100	62/100	100

Income level	Own a smartphone		Row Total
	No	Yes	
High	10.00%	90.00%	20
Medium	40.91%	59.09%	66
Low	64.29%	35.71%	14
Column Total	38.00%	62.00%	100



Now let us look at the second example. In the second example I do the same thing I know the row totals were 20 for a high income group, 2 people did not own a phone so 2 by 20 which is a 10%, 18 by 20 which is 90%. So, the way I can articulate this is of the high income group or 90% of the high income group people own a phone whereas 10% of the high income group do not own a phone.

In the medium income group 40% or 41% of medium income group do not own a phone 60% own a phone or 59% own a phone and this is 41%, this is 59% own a phone whereas, in the low income group I have a whopping 64% who do not own phones and a 36% who own phones. In total I have 38 people who do not own 38% who do not own and 62% who own a phone. So, these are what we refer to as row relative frequencies.

(Refer Slide Time: 08:04)



Column relative frequencies

- ▶ What proportion of total participants are female?
- ▶ What proportion of smart phone owners are females?

Gender	Own a smartphone		Row total
	No	Yes	
Female	10	34	44
Male	14	42	56
Column total	24	76	100

Column relative frequency: Divide each cell frequency in a column by its column total.



Similarly to the low relative frequency I have what is called column relative frequencies. Now what are the type of questions we are expected to answer here. Let us go back to our contingency table here. So, now I want to know what is the proportion of total participants who were female? I have 100 participants the proportion is 44 by 100 so I have 44% female and 56% male participants.

Now of the people who own smart phones. So, I have 76 people who own smart phone I want to know among this what is the proportion of females among the smart phone owners. The answer to this question is total number of smart phone owners are 76 of which 34 of them are female. So, the proportion of people among the smart phone owners who are female are 34 by 76, proportion of male people who are owners are 42 by 76, proportion of female non owners are 10 by 24, proportion of male non owners are 14 by 24 these values 10 by 24, 14 by 24, 34 by 76 and 42 by 76 are what we refer to as a column relative frequency.

How do we obtain the column relative frequency? We divide each frequency by their respective column totals we get the column relative frequency.

(Refer Slide Time: 10:04)



Example 1: Column relative frequency

Gender	Own a smartphone		Row total
	No	Yes	
Female	10/24	34/76	44/100
Male	14/24	42/76	56/100
Column total	24	76	100

Gender	Own a smartphone		Row Total
	No	Yes	
Female	41.67%	44.74%	44.00%
Male	58.33%	55.26%	56.00%
Column Total	24	76	100



So, you can see the relative frequency is 41% of non owners are female, whereas 58% of non owners are male whereas when it comes to owning a cellphone 44% are female and about 55% are male. Totally 44% female and 56% are male. So, these are the column relative frequencies. Now let us look at the row and column relative frequencies for the second example which was the income versus the ownership of a smart phone.

(Refer Slide Time: 10:51)

Statistics for Data Science - I
↳ Association between categorical variables
↳ Relative frequencies

Example 2: Column relative frequency



	Own a smartphone		
Income level	No	Yes	Row total
High	2/38	18/62	20/100
Medium	27/38	39/62	66/100
Low	9/38	5/62	14/100
Column total	38	62	100

	Own a smartphone		
Income level	No	Yes	Row Total
High	5.26%	29.03%	20.00%
Medium	71.05%	62.90%	66.00%
Low	23.68%	8.06%	14.00%
Column Total	38	62	100

Now when you look at the column relative frequency for the second example you see that 20% high income, 66% medium income and 14 % low income group. Now among the owners you can see that among the people who own a phone, 29% of the people who own a phone are from the high income, 63 are from medium and a low 8% are from the low income groups. Now when it comes to the people who do not own a phone you see only 5% are from the high income group, 71% are from the medium and 24% from the low income group. So, this is how you compute the relative frequency.

(Refer Slide Time: 11:53)



Section summary

- ▶ Concept of relative frequency: row relative frequency and column relative frequency.



We have introduced a concept of a relative frequency. In particular we have seen how to compute the row relative frequency and the column relative frequency. We now we will see how to use the concept of a row relative frequency and a column relative frequency to answer questions about association between variables.

(Refer Slide Time: 12:22)



Association between two variables

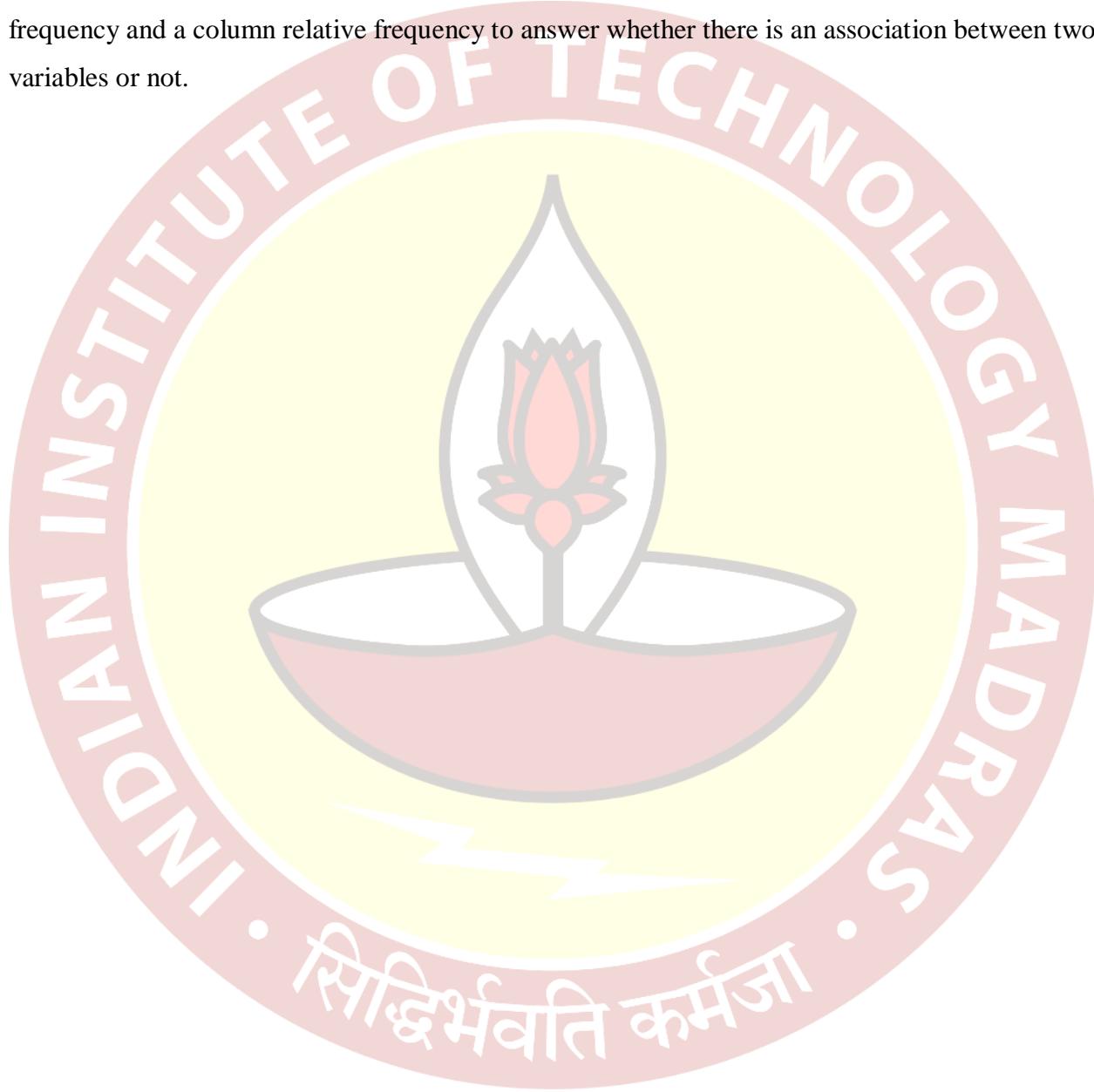
1. Contingency table - Summarizes
2. Row relative freq - $\frac{\text{Cell freq}}{\text{Row total}}$
3. Col. relative freq - $\frac{\text{Cell freq}}{\text{Col. total}}$



So, now we address the questions which we started this module with or this lecture with by wanting to answer whether there is an association between two variables. We first introduce how to set up what we call a contingency table. A contingency table basically summarizes your bivariate data

then we introduced a notion of both a row relative frequency, to compute the row relative frequency you measured each or you divided each cell frequency by its row total.

And then we also introduced a notion of a column relative frequency where again I divided each cell frequency by its column total. Now we will see how to use this concept of a row relative frequency and a column relative frequency to answer whether there is an association between two variables or not.



(Refer Slide Time: 13:35)

Statistics for Data Science -I
↳ Association between categorical variables
↳ Association between variables



Association between two variables

- ▶ What do we mean by stating two variables are associated?
Knowing information about one variable provides information about the other variable.
- ▶ To determine if two categorical variables are associated, we use the notion of relative row frequencies and relative column frequencies described earlier.



So, what do we mean by saying or stating the two variables are associated. In other words, we want to know that whether information about one variable provides some information about another variable. So, when we are seeking to answer the question whether two variables are associated actually what we are seeking to answer is whether if I have information about a particular variable whether it actually gives me something or tells me something about the other variable. So, to determine whether two categorical variables are associated we will now show how we use the notion of relative row frequencies and relative column frequencies.

(Refer Slide Time: 14:32)



Association between two variables

- ▶ If the row relative frequencies (the column relative frequencies) are the same for all rows (columns) then we say that the two variables are not associated with each other.
- ▶ If the row relative frequencies (the column relative frequencies) are different for some rows (some columns) then we say that the two variables are associated with each other.



So, let us look at our relative row or column frequencies. We already know how to compute these frequencies. If the row relative frequency or the column relative frequencies are the same for all rows. I repeat, if the row relative frequency or the column relative frequencies are the same for all rows or columns, then we say the two variables are not associated with each other.

If the row relative frequency or the column relative frequencies are different for some rows then we say that the two variables are associated with each other. So, if the row relative or the column relative are same we say they are not associated. If the row relative frequency or the column related frequency are different, then we say they are associated with each other.

(Refer Slide Time: 15:45)



Example 1: Association between two variables

- If the row relative frequencies (the column relative frequencies) are the same for all rows (columns) then we say that the two variables are not associated with each other.

Gender	Own a smartphone		Row total
	No	Yes	
Female	22.73%	77.27%	44
Male	25.00%	75.00%	56
Column total	24.00%	76.00%	100

Gender	Own a smartphone		Row Total
	No	Yes	
Female	41.67%	44.74%	44.00%
Male	58.33%	55.26%	56.00%
Column Total	24	76	100

Gender and smartphone ownership are not associated



So, let us go back to our examples to and apply this rule which compares the row relative frequency and the column relative frequency. So, let us look at this example in this example where I plotted or I tabulated the gender versus ownership of a phone you can see that when it comes to a ownership of a phone 24% of a total population did not own a phone, 76% owned a phone.

Now if I am looking at the pattern within the gender you see again 23% of the females did not own a phone and about 77% of the females owned a phone. When it comes to male again I see about 25% did not own a phone and 75% owned a phone. So, you see that the ownership pattern which is 24% not owning and 76% owning a phone is consistent with both the female subgroup and the male subgroup.

You do not see any inconsistencies that is both when you look at females also you see that about 23% are not owning and 77% are owning. In the male also we see about 25% are not owning and 75% are owning. So in general, the ownership pattern does not change depending on the gender. Let us look at the column frequencies. Again if you look at the column frequencies I had 44 % female and 56% male.

Now if you look at only owners of the 76 people again I see about about 45% are female and 55% are males which is almost the same as my total gender diversity. This same percentage among people who do not own a phone again I have about 42% females and 58% males. So, the gender

diversity also among owners of a phone and not owners of a phone is also the same that is 44% and around 44% and 56%.

So, both the row relative frequencies and the column relative frequencies are the same for all the rows and columns. Hence, I can say that both my gender and smart phone are not associated with each other which is consistent with the definition earlier. Now let us look at the second example.

(Refer Slide Time: 19:02)

Statistics for Data Science -II
↳ Association between categorical variables
↳ Association between variables

Example 2: Association between two variables

► If the row relative frequencies (the column relative frequencies) are different for some rows (some columns) then we say that the two variables are associated with each other.

Own a smartphone		Row Total	
Income level	No		Yes
High	10.00%	90.00%	20
Medium	40.91%	59.09%	66
Low	64.29%	35.71%	14
Column Total	38.00%	62.00%	100

Own a smartphone		Row Total	
Income level	No		Yes
High	5.26%	29.03%	20.00%
Medium	71.05%	62.90%	66.00%
Low	23.68%	8.06%	14.00%
Column Total	38	62	100

Income and smartphone ownership are associated



When I look at the second example I plot both the row relative frequencies, what are my row relative frequencies here. So, you have here what were the two variables; the income level and whether you own a phone or not. Again let us look at it in the first case I know 38% do not own a phone and 62% own a phone, but when you look at the high income group you see that 90% own a phone and 10% do not own a phone whereas, in the low income group, 65% do not own a phone and 35% own a phone.

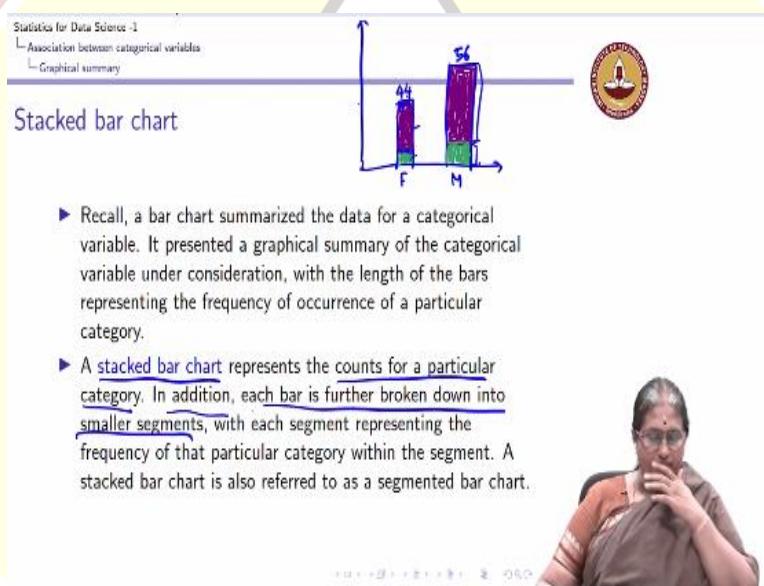
So, you can see that whether you own a phone or not is playing the percentages of ownership of a phone actually is different for the high income group and the low income group. The row relative frequencies are not the same among the categories. If you look at the column relative frequencies again I have a distribution of 20%, 66% and 14% in my high, medium and low income groups.

If I just look at the owners I have 62% who are owners of which I have about 30% who are coming from the high income group and only a 8% who come from the low income group. Among the non

owners I have a 5% who are from the high income group and a whopping 24% who are from the low income group. This is not consistent with my total distribution of my income categories.

Since the row frequencies and the columns frequencies are different among the rows and the column, I can say that the income and smart phone ownership are actually associated with each other which is very intuitive. You would expect the ownership of a phone to actually be associated with your income level whereas the ownership of a phone need not be associated with whether you are a female or a male and we have seen both the examples which actually have demonstrated this phenomena.

(Refer Slide Time: 21:50)



So, we have seen how we can use the concept of a relative frequency to decide whether two variables are associated or not. We again recall we said that if the row relative frequencies or the column relative frequencies are same for all the rows and columns we say two variables are not associated with each other. If they are same if they are not same or if they are different we say that they are associated with each other and we demonstrated this through the two examples which we have been discussing.

Now, how do I graphically show this result? So, again let us go back to our examples here see that we have a contingency table which is given here. Recall, when we wanted to summarize a single categorical variable we used what was called a bar chart. Now, I want to see how these two

variables behave with each other. So, for that what I do is I construct what is called a stacked bar chart or sometimes it is also referred to as a segmented bar chart. So, a bar chart summarize the data for a categorical variable where the length of the bars were representing the frequency of occurrence of a particular category.

This is what a bar chart does. Now a stacked bar chart represents the counts for a particular category in addition each bar is further broken down to smaller segments. Now let us illustrate what we mean by this? Now if we are looking at a bar chart for the first example I had two categories the female category and the male category. Among the female I had 44 females and 56 males and you see that this is my bar chart.

Now what is a stacked bar chart? Now if I want the second category to be super imposed on this. What do I mean by this? Again you go back to your contingency table. You see that out of 44 in the first example I have out of 44 people I have 77.27 who own a phone that was this was 30 actually among the 44 people I have 77. So, this was 34 and this was 10 this was how I had it and this was a 14, 46.

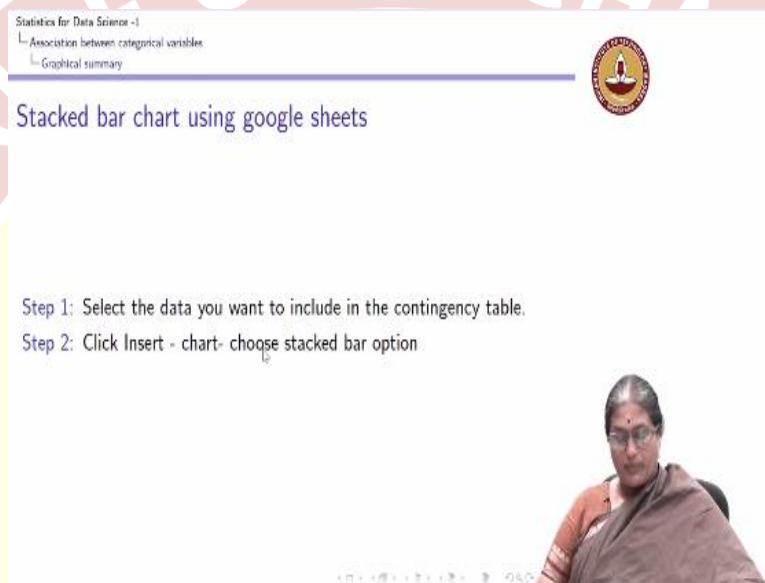
This was not 46 this was a 14 and this was a 42. So, if I now given this bar chart I know that 77% so I have constructed this bar chart. In this I want to know what percentage of the female own a phone so what I can do in that case is I can look at a percentage let me look at another color. So, I have about 77% of this who own a phone approximately this is about 77%.

So, of this 44 I have about 77% similarly here I have the same percentage who own the phone here so this purple shaded area represents the number of people or the proportion of people within each category who own a phone. I have about 77% who own a phone and this green shaded area represents the proportion of people among females who do not own a phone and this green area here represents the proportion of males who do not own a phone.

So, the difference between a staked bar chart and a bar chart is, when I have only the female and male category I could have. So, this entire bar represented the count of female and this entire bar represented which was 56 represented the count of female whereas a stacked bar chart in addition to the so what you can see that in addition to the count of a particular category it breaks it down into smaller segments.

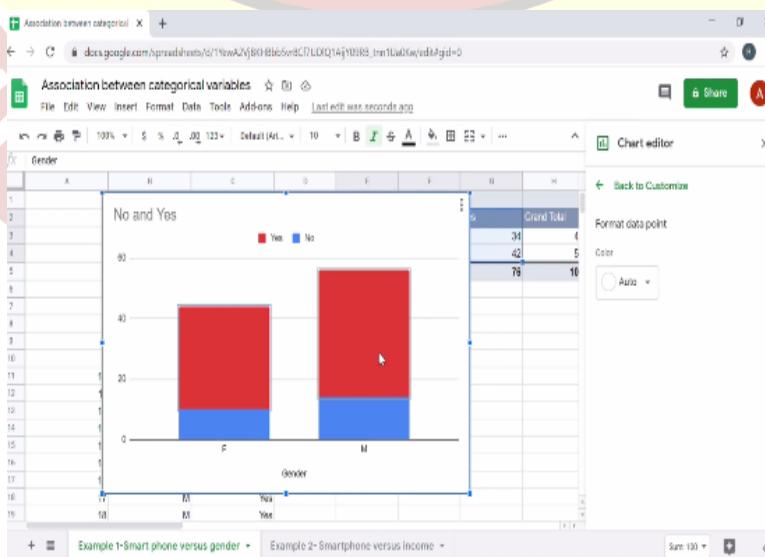
So, I broken down this entire of 44 into smaller segment. Here again two segments where this segment represents the owners of or the female owners and the green segments represents the female non owners. Similarly, this segment represents the male owners and this segment represents the male non owners of the graph. So, since you have the segmented bars it is called a segmented bar chart or it is also known as a stacked bar chart.

(Refer Slide Time: 27:53)



How do we construct a stacked bar chart using Google sheet.

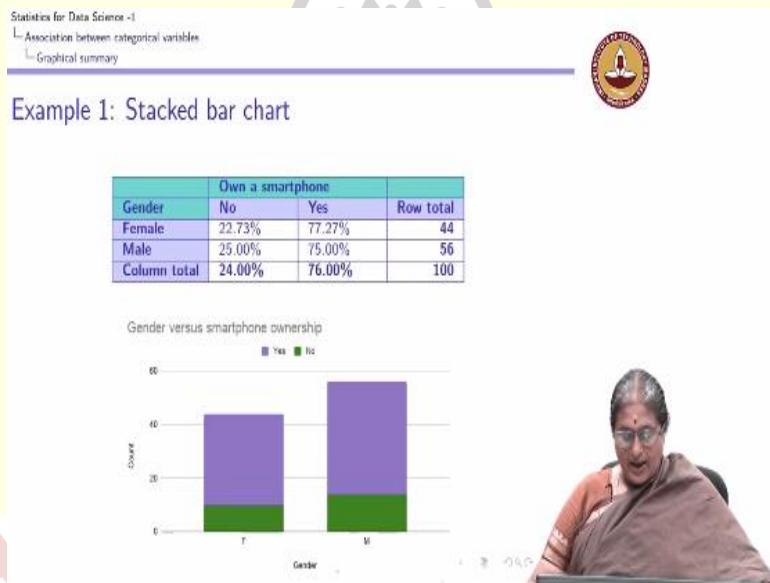
(Refer Slide Time: 27:55)



So, now we go back to the contingency table which we had already constructed. I select the data for which I have the contingency table I am selecting the gender and the no, yes I go to insert chart in an insert chart I am looking at a stacked column chart. You can see that I am looking at a stacked column chart. So, you can see that this is the stacked column chart of which you can also see that I have two genders. Here I have a female I have a male.

This is the gender I have and within the female I have 44 within the male I have 76. 77% of the female gender is actually owning a phone the red color indicates yes the blue color indicates no. Again close to 77% of male own a phone again yes. So, within the male bar I have indicated how many own a phone and how many do not own a phone.

(Refer Slide Time: 29:21)

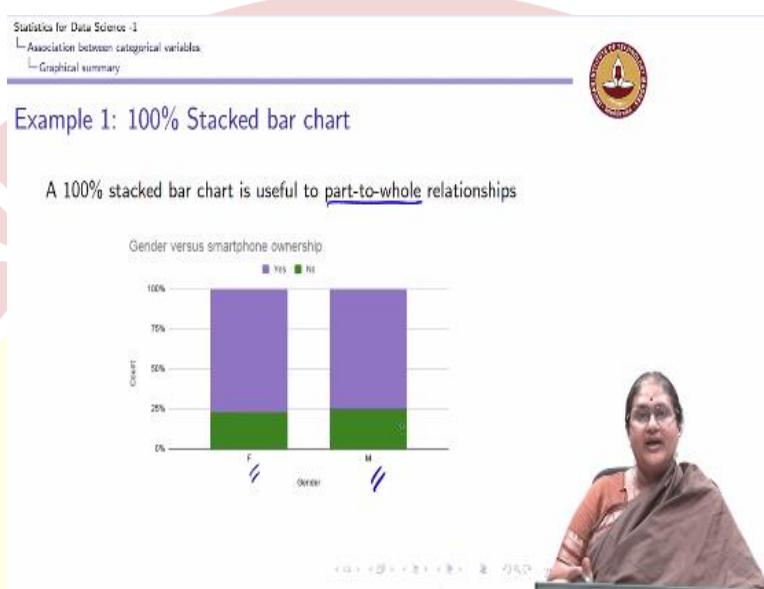


So, you can see that a stacked bar chart is a good way of summarizing the ownership in a graphical way. Now you can go back to your Google sheet and you can see that this stacked bar chart give me the actual counts, but suppose I am interested and this was what we referred to as a standard stacked bar chart. So, you can go to a chart style go to setup I have a stacked column chart.

Under stacking you can see that I have listed the standard option, but if I click on what I call a 100% stacked bar chart. What a 100% stacked bar chart gives me is you can see that of if I consider it does not give me the actual counts of a female and male, but what a 100% stacked bar chart gives me is the proportion of females who own a phone to the proportion who do not own.

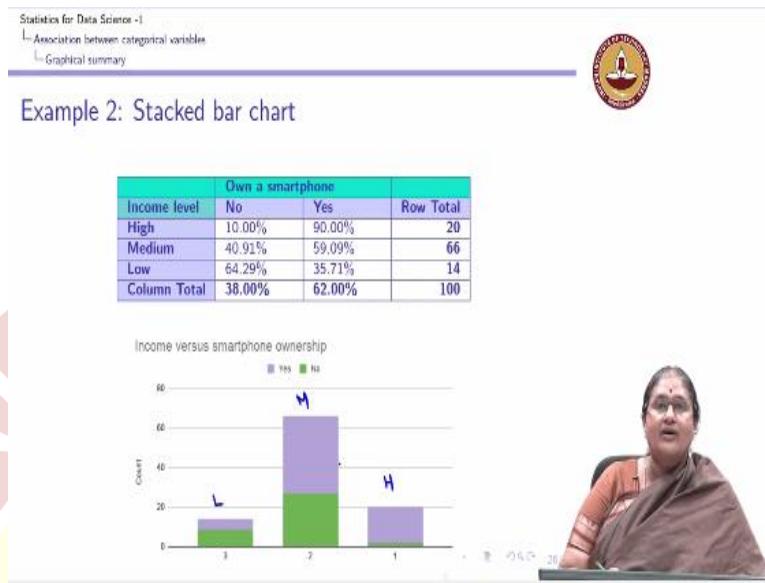
And similarly the proportion of males who own a phone to proportion who do not own. Now, where is a 100% stacked bar chart useful to me. Now suppose I go to a 100% stacked bar chart. So, where is a 100% stacked bar chart useful.

(Refer Slide Time: 31:07)



You can see that when I am actually not interested in knowing the count of each category, but I am interested in knowing about a part-to-whole relationship or the proportional relationships I can use what is a 100% stacked bar chart. Here you can see it very visually it is showing me that the distribution or ownership of a phone to not having a phone for the category female and male is almost the same you do not see any sizable difference between the ownership pattern and gender.

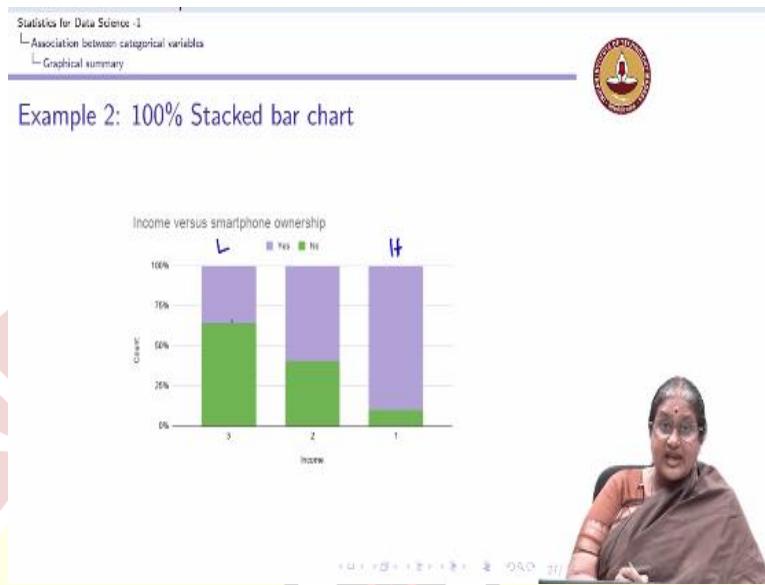
(Refer Slide Time: 31:50)



For the second example when you look at a stacked bar chart. Now if you look at this chart I have the category 1. 1 you recall was associated with the high income group, 2 was associated with the medium income group, 3 was associated with the low income group again their order I either have to maintain a low medium high order or a high medium low order. This you have to maintain the order you cannot have a low, high, medium order because it is an ordinal data maintain the order in which you represent the variables.

Now if you look at this high income group yes is the purple, green is the no. You can see that among the high income group you have more people who own the phone to people who do not own the phone. In the low income group you have more people who do not own the phone to people who own the phone and in the medium income group it is almost you have equal number of people who own a phone to do not own a phone. Graphically this is very clear.

(Refer Slide Time: 33:02)



Now, how does a 100% stacked bar chart look for this example. Now if you look at a 100% stacked bar chart where I am not interested in the actual counts, but I am interested in looking at how the 100% stacked bar chart looks. You can see this very clearly in the high income group I have a lot of people who own the phone. In the low income group I have this green is higher than the purple.

Green is the number of people who do not own the phone whereas for medium I have equal number of the proportion of people who own the phone is equal to the proportion of people who do not own a phone. So, you can see when you are not interested in the actual counts, but you are interested in comparing these groups with each other to tell you a story a 100% stacked bar chart is very useful.

And this story is since you do not see a varying pattern we can reaffirm what we saw from the column and row relative frequency that income and ownership are associated with each other. Whereas, when we looked at gender versus ownership they were almost the same gender and ownership are not associated with each other.

(Refer Slide Time: 34:36)

Statistics for Data Science - I
└ Association between categorical variables
 └ Graphical summary

Section summary

▶ Understand whether two categorical variables are associated using the concept of relative frequencies.
▶ Graphical summary of association using stacked bar chart.

So, at the end of this section you should know, you should be able to use the concept of relative frequency to tell whether two variables are associated with each other. You further validated through a graphical summary. This graphical summary is what we referred to as the stacked bar chart.

Statistics for Data Science 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture 4.4
Association between Two Numerical Variables: Scatterplot

(Refer Slide Time: 00:16)

Statistics for Data Science -1
└─ Association between numerical variables
 └─ Introduction

Introduction

- ▶ To understand the association between two numerical variables.
- ▶ Learn how to construct scatter plots and interpret association in scatter plots.
- ▶ Summarize association with a line.]
- ▶ Correlation matrix ↗



The next part of this lecture, we are going to understand how we describe the association between 2 numerical variables. What are we going to do in this case; we first introduced what is a scatter plot. And then we try and interpret the association between the 2 variables using the scatter plot. Recall, we interpreted the association between 2 categorical variables using the notion of a contingency table.

And here we are going to use a notion of a scatter plot. And then further, we will just briefly introduce how we summarize this association through a concept of a line. And then we will also introduce the notion of a correlation matrix so that this notion can be extended to understand association between more than 2 variables.

(Refer Slide Time: 01:14)

Statistics for Data Science -I
└ Association between numerical variables
 └ Scatter plots

Scatter plot

We use a scatterplot to look for association between numerical variables.

Definition

A scatter plot is a graph that displays pairs of values as points on a two-dimensional plane.

- ▶ To decide which variable to put on the x-axis and which to put on the y-axis, display the variable you would like to explain along the y-axis (referred as response variable) and the variable which explains on x-axis (referred as explanatory variable).

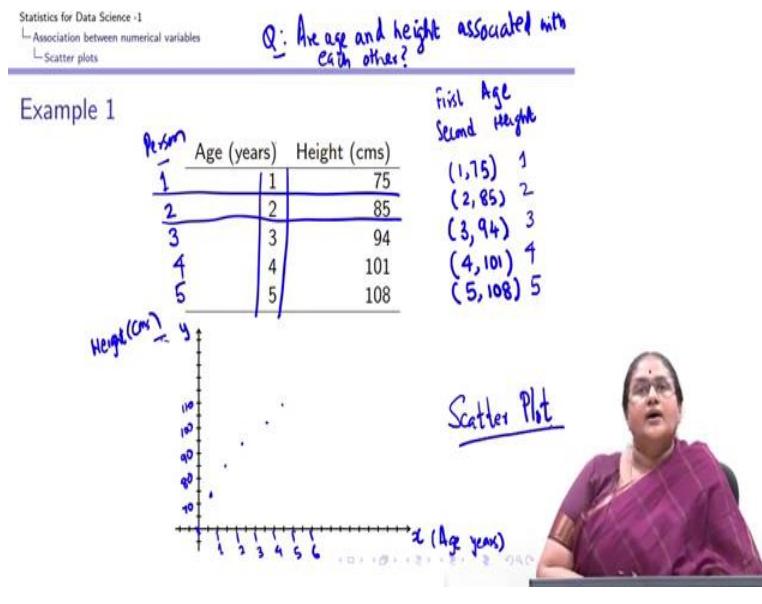


Navigation icons: back, forward, search, etc.

So what is a scatter plot? A scatter plot is defined as a graph that displays pairs of values as points on a 2-dimensional plane. So, what do we mean by this?



(Refer Slide Time: 01:33)



Scatter plot

We use a scatterplot to look for association between numerical variables.

Definition

A scatter plot is a graph that displays pairs of values as points on a two-dimensional plane.

- To decide which variable to put on the x-axis and which to put on the y-axis, display the variable you would like to explain along the y-axis (referred as response variable) and the variable which explains on x-axis (referred as explanatory variable).



So, let us start with a very simple example. Suppose I am looking at 2 variables, what are the 2 variables I am looking at? The first variable I am looking at is age. And the second variable I am looking at is height. The people are the same. So I have an observation here. So, this is my first or this is the person, okay, the first person is of age 1 and 75 centimeters; person 2, person 3, person 4, person 5 and I am just looking at the age of this person.

So what this means, and the way I can interpret is, the person's 1 age is 1 and 75 centimeters, I can write this as an order pair 1, 75. For the person 2 age is 2 and height is 85 centimeters, age is measured in years, height is measured in centimeters, person 3 is 3 and 94 centimeters, person 4

is 4 and 101 and person 5 is 5 and 108 centimeters, okay. So, I have 2 variables, both of the variables are numerical.

And I am measuring these 2 variables age in units of years and height in units of centimeters. And the question we are asking here is are age and height associated with each other? This is the question we are asking. This is the question we are asking. Now to answer this question, we first try and plot it as a scatter plot on a 2-dimensional plane. Now when I am plotting it on a 2 dimensional plane, all of us know that this axis is called the x axis and this axis is called the y axis.

So the first thing which we need to decide is which of these 2 variables would go on the x axis and which variable would go on the y axis. The rule of thumb is when I want to understand the association between any 2 variables typically I might want to know whether one variable is being explained by the other variable.

In this example, I would want to know whether as a person grows older, I want to know what is the height or what is the association with the height does the height of an individual increase or does it decrease or is there any association at all? So the variable that goes on to your x axis that is also referred to as the explanatory variable.

The variable that you are using to explain and the variable you would like to explain, which is height in this example, is also referred to as a response variable that is one my Y axis. So, I can write my response variable on my Y axis. In this case, it is the height which is measured in centimeters. And my explanatory variable is age which is again measured in years. So, once I have the age and height written then I will go and start writing on my x axis.

Remember, we are again talking about numerical data, there is an order to it, so, I have a start with the 0 yet I have a 1, I have a 2, I have a 3, I have a 4, I have a 5, I have a 6. I can start putting my data on my x axis and what is the data on my x axis? It is the age in years that is what is given to me. Now, when I look at the height, I can start and remember I can put a break here because I wants to start with 70 centimeters, I have 80, 90, 100, 110.

So, first age 75. So, this would be the point which is associated with my first point with 2 I have 85 which is this point, with 3 I have 94 which is this point, with 4 I have 101, which is again this

point and with 5 I have 108, which is this point. So, in effect what we have done is, we have actually plotted the data that is a pair of values.

What are the pairs of values we have here? This is the first pair this is the second pair, this is the third pair, this is a fourth pair and this is the fifth pair, I have these 5 pairs of values, which I have represented as points on my 2-dimensional plane. So, this plot which I have constructed here is what we refer to as a scatter plot.

(Refer Slide Time: 07:20)



A real estate agent collected the prices of different sizes of homes. He wanted to see what was the relationship between the price of a home and size of a home. In particular, he wanted to know if the prices of homes increased linearly with the size or in any other way? To answer the question, he collected data on 15 homes. The data he recorded was

1. Size of a home measured in 1000 of square feet.
2. Price of a home measured in lakh of rupees.

Explanatory : Size
Response : Price

Navigation icons: back, forward, search, etc.

So, the scatter plot is an extremely powerful graph, which generally is just the scatter or it is a display of pairs of values of my variables. Now, let us look at another example. So, a real estate agent has collected the prices of different size of homes. So, what the real estate agent has done? He has gone he has collected 1, 2, 3 up to 15 homes. So, the real estate agent has collected on every home the size of the house, this is measured in thousands of square feet and the price of the house in lakhs of rupees. This is what I have as my data.

So what is it I am seeking? The seeking the question I am seeking an answer to is whether there is a relationship between the price of a home and size of home. Now in this example, you very clearly see that I want to see whether the prices vary according to the sizes. So, the natural explanatory variable in this case in this example is going to be the size of a house, whereas my response variable in this case is going to be the price of a house.

So, whenever you want to understand the association between 2 variables and you are interested in plotting or coming up with a graphical display in terms of a scatter plot. The first step is to recognize what is your explanatory variable and what is your response variable. So, next he wanted to know whether the prices of homes increase linearly with this size. We will come to answer this question in some time, but even before that, we want to see what is the data is recorded.

(Refer Slide Time: 09:29)

Statistics for Data Science -1
↳ Association between numerical variables
↳ Scatter plots

Housing data

	Size (1000 Square feet)	Price (INR Lakhs)
1	0.8	68
2	1	81
3	1.1	72
4	1.3	91
5	1.6	87
6	1.8	56
7	2.3	83
8	2.3	112
9	2.5	93
10	2.5	98
11	2.7	136
12	3.1	109
13	3.1	122
14	3.2	159
15	3.4	170

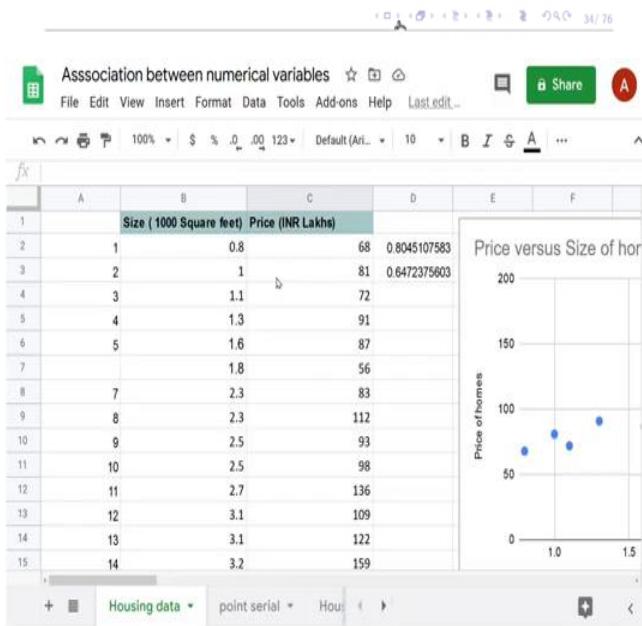
So, this is my data. So, I have a different data given here. So with this data, what we do is if I am just going to so I have 15 data points and I have these are the lakhs INR, lakhs or 68 lakhs is 800 square feet house is costing me 68 lakhs, 1000 square feet is 81 lakhs, 1100 square feet is 72 lakhs and 3400 square feet house is 170 lakhs. So what I do now I just so now I have, I need to plot a scatter plot. So on my x axis I am going to take the size. On the y axis, I take the price of the house plot a scatter plot, how do I plot a scatter plot.

(Refer Slide Time: 10:16)

Statistics for Data Science -I
└ Association between numerical variables
└ Scatter plots

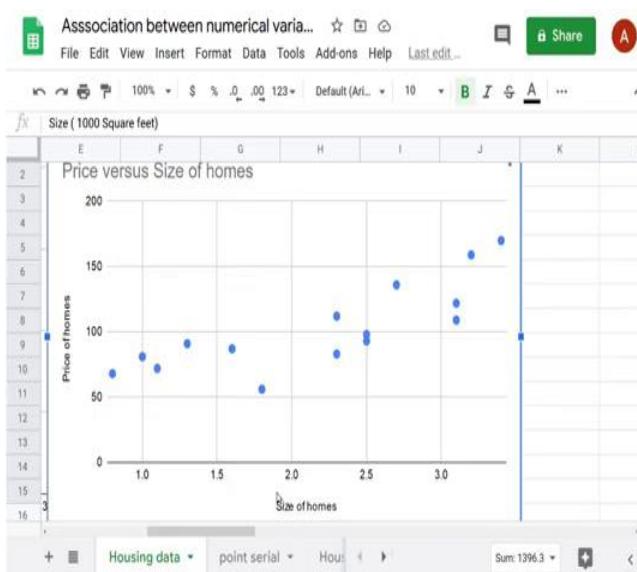
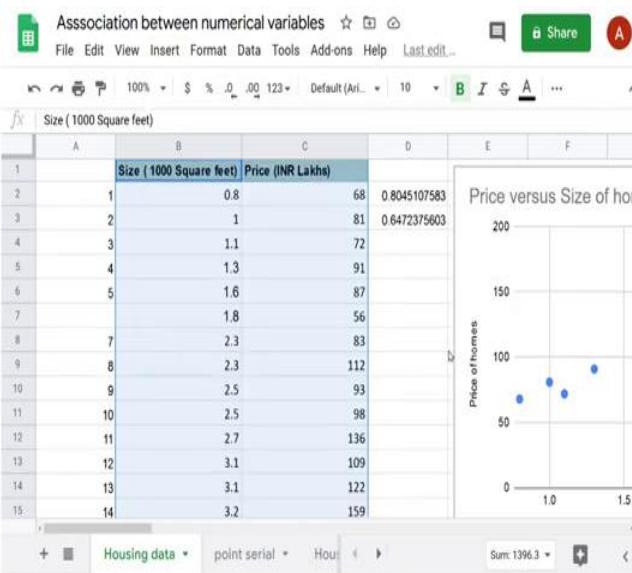
Scatter plot using google sheets

- Step 1: Highlight data you want to plot
- Step 2: Insert - chart- choose scatter chart
- Step 3: Under X-axis tab, choose your explanatory variable.
- Step 4: Under series tab, the response variable.
- Step 5: Label the title of the chart, axes appropriately.



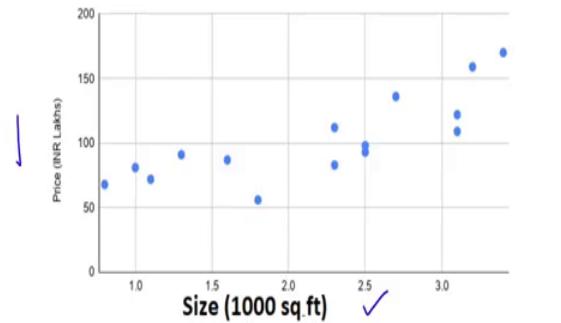
So let us go back to our Google Sheets. This is the data given I have the same data given in my Google Sheet. So you can see that this is the same data I have 0.8, 0.1, 1.1 so I have 68, 81, 72 this is in square feet, and this is in lakhs of rupees. So the first thing I want to ask is can a plot a scatter plot, the Google Sheet gives me an easy way to plot a scatter plot. Let us go about and plot a scatter plot using a Google Sheet.

(Refer Slide Time: 10:51)



साक्षिर्भवति कर्मजा

Scatter plot

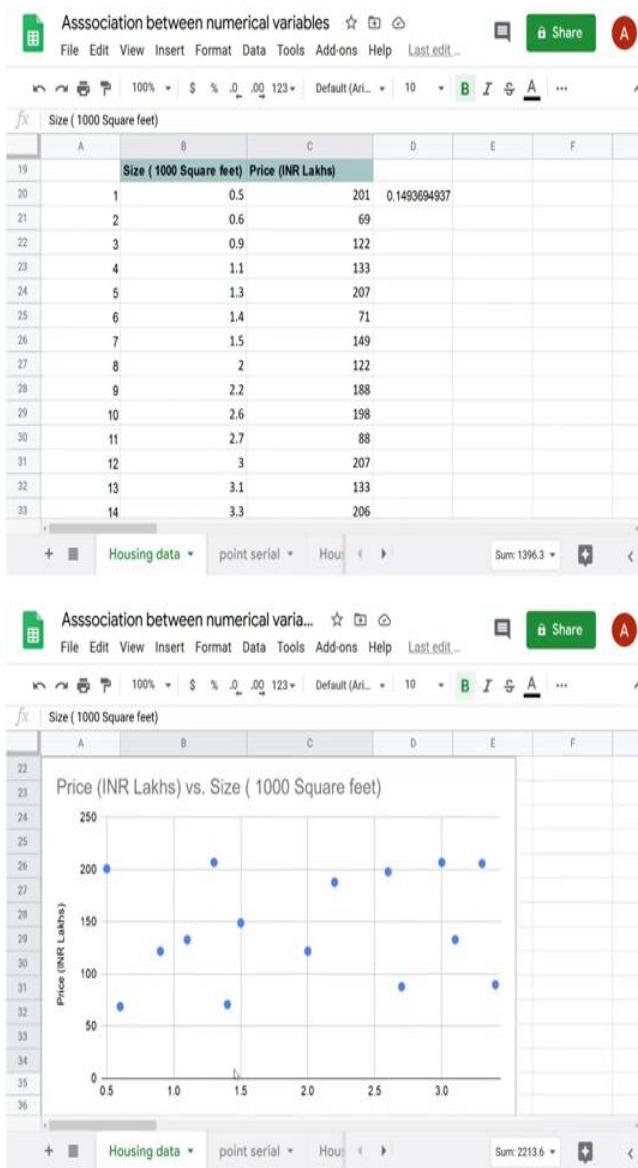


So the first step is highlight the data you want to plot as a scatter plot. And I have highlighted that data insert chart. So if I go to insert and go for a chart, you can see that I have a scatter chart, which has appeared. And this is what it is under the x axis tab, the x axis (I have) it chooses the square feet, and it has automatically chosen in this case, because that was my first column available.

And in under the series tab, it gives me what is the response variable, which is again price in INR lakhs, and I can again, label, title and access appropriately. So you can see that I have a nice scatter plot between the size and the price which I got from my Google Sheet. So this is the scatter plot we have got, again, I have size on my x axis, and price on my Y axis. So this tells us how to construct a scatter plot.



(Refer Slide Time: 12:05)

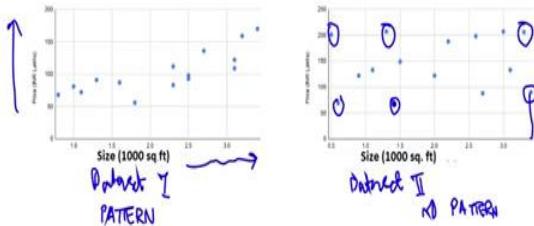


Now once given a scatter plot, let us look at another example. For example, if I go and look at this other example I have here, so the same price so I remove this, let me remove this graph, because I have already I delete this chart. Let me go to another example where I am again, talking about same the price and the size versus the price. This is the data I have, but if I am plotting a scatter plot between this data, you can notice a scatter plot of this kind, okay.

(Refer Slide Time: 12:47)

Visual test for association

- Do we see a pattern in the scatter plot?
 - In other words, if I know about the x-value, can I use it to say something about the y-value or guess y-value?



36 / 76

So, let us copy the scatter plot. So, what we do is this is the scatter plot again 15 homes I have a scatter plot here. So this graph both the graphs actually plot size versus price. This is for my first data set which I call data set 1 and this is for the second data set which I refer to as data set 2, okay. So now, if you look at this in data set 1, I can observe that as the sizes of the homes are increasing, the prices are also exhibiting some sort of a trend or it increases.

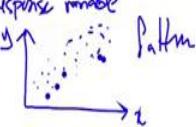
Whereas in this I have smaller homes for example, if I look at this size of a home, I have bought a lower priced house and higher priced house and I have larger homes also this is a larger home because the size is a larger size, this size also I have a lower price and a higher price. A midsize home also has a lower price and a larger price. In other words, this data set 1 has a pattern which I can in some sense explain whereas in data set 2 I do not see any clear pattern.

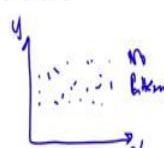
So, the first thing which we want to understand is, remember, we wanted to know whether a scatter plot can actually help us understand association between numerical variables. So, we can see that from these 2 examples in the case of 1 example I can actually see a pattern whereas in the other case, I do not see any clear pattern. So the visual test for association tells us whether you can see a pattern or not see a pattern.

(Refer Slide Time: 14:54)

Statistics for Data Science -I
└ Association between numerical variables
└ Association in scatterplots

Section summary

① Identify ✓ Explanatory variable
✓ Response variable
② 

✓ Draw a scatter plot
↳ Notion of explanatory variable and response variable.
✗ Visual test for association 

37/76

So, what we have learned so far as a first given data set identify what is your explanatory variable and what is your response variable. Given this, let your explanatory variable be on the x axis, response variable on the y axis, draw a scatter plot.

Once you have a scatter plot then I could have a scatter plot which is of this kind where I see a pattern or I could have a scatter plot which is of this kind where I really do not see any clear pattern between my x and y. This is referred to as a visual test for association and how do we actually draw conclusions from this visual test is what we are going to see next.



Statistics for Data Science 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture 4.5
Association between Two Numerical Variables: Describing Association

(Refer Slide Time: 00:15)



When describing association between variables in a scatter plot, there are four key questions¹ that need to be answered

- 1.. Direction: Does the pattern trend up, down, or both? ✓
2. Curvature: Does the pattern appear to be linear or does it curve? ✓
3. Variation: Are the points tightly clustered along the pattern?
4. Outliers: Did you find something unexpected?



¹Stine, Robert, and Dean Foster. Statistics for Business: Decision Making and. Addison-Wesley, 2011.

So how do we describe the association between the 2 variables using a scatter plot? So when we are describing association between 2 variables, there are 4 key questions that I need to answer. The first question is, is there a direction? What do I mean by a direction? Does the pattern trend up or down or does it exhibit some sort of a trend? Is it linear or does it curve? The third is, are the points tightly clustered around the pattern or are they spread? Do I find anything unexpected? We will look at each one of these questions in detail now.



(Refer Slide Time: 01:05)

Statistics for Data Science -I
└ Association between numerical variables
 └ Describing association

Describing association: Direction

Does the pattern trend up, down, or both?

Price (INR Lakh) vs. Size (100 Square feet)

as size increases
price increases

Navigation icons: back, forward, search, etc.

Statistics for Data Science -I
└ Association between numerical variables
 └ Describing association

Describing association: Direction

Does the pattern trend up, down, or both?

Price (INR Lakh) vs. Size (100 Square feet)

Up

Price vs. Age of a car (years)

E.g.: Age of a car
Rep = Price fair

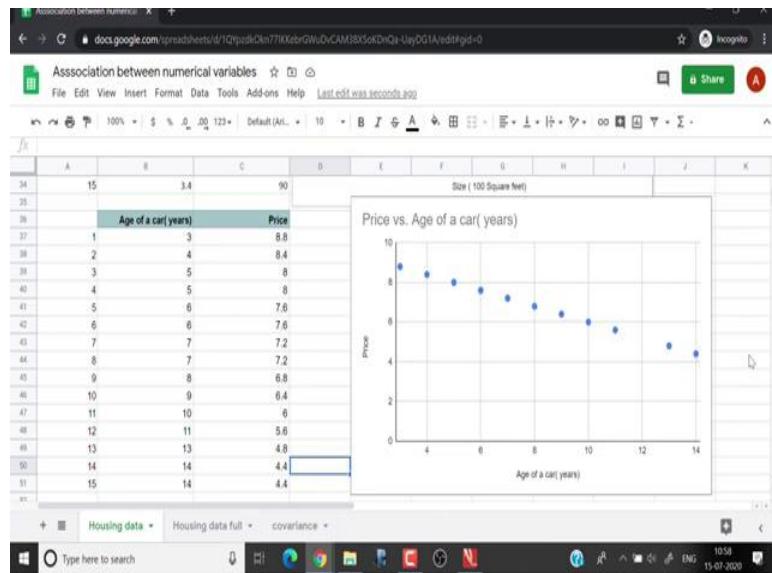
Navigation icons: back, forward, search, etc.

So, the first question we want to know is, does the pattern trend up or down? Let us look at an example again, if you look, go back to this example where I had plotted the size of a whole house on my x axis, and I wanted to know the price versus size, we see that there is a pattern where I can easily say that as the sizes of my homes increase.

As size increases, price increases. I can see there is an upward trend, so the trend is up, okay. So, let us look at another example. We know that as car ages, so here instead of looking at an age of a person, my explanatory variable is age of a car. And my response variable is the price of the

car. So what I mean by this is as a car becomes older or the older the car is the price I am going to get for that car again reduces.

(Refer Slide Time: 02:32)



So let us look at a data here. So you can look at the data for a 3 year old car if I am getting the 8.8 lakhs again prices in lakhs of rupees for a 4 year old car I might get 8.4, for a 5 I might get 8, for 6 I would get 7.6 lakhs. So you can see that as the car becomes older my prices showing is coming down.



(Refer Slide Time: 02:58)

Statistics for Data Science -I
└ Association between numerical variables
 └ Describing association

Describing association: Direction

Does the pattern trend up, down, or both?

Price (INR Lakh) vs. Size (100 Square feet)

Up

Price vs. Age of a car(years)

as Age increases
Price decreases

Fxp: Age of a car
Rxp= Price per car

388 / 58

Statistics for Data Science -I
└ Association between numerical variables
 └ Describing association

Describing association: Direction

Does the pattern trend up, down, or both?

Price (INR Lakh) vs. Size (100 Square feet)

Up

Price vs. Age of a car(years)

Down

388 / 58

And that is what is shown by the scatter plot here. So what is the question or what is the pattern we see here? As the age increases, I see the price decreases. In this case I saw as size increases price increases. So here I can describe my pattern to have a decreasing or down trend. So the first thing which we need to see when we look at our scatter plot says whether it is showing an upward direction or a downward direction.

(Refer Slide Time: 03:45)

Statistics for Data Science -I
 └ Association between numerical variables
 └ Describing association

Describing association: Curvature

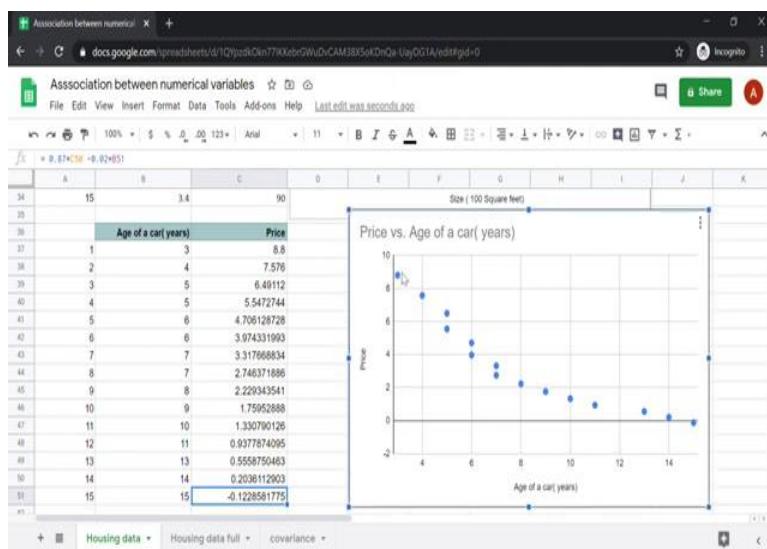
Does the pattern appear to be linear or does it curve?

Price (INR Lakhs) vs. Size (100 Square feet)

Price vs. Age of a car (years)

Price (INR Lakhs) vs. Size (100 Square feet)

Price vs. Age of a car (years)



The next thing is, does the pattern appear to be linear or does it curve? Is it a linear pattern or is it a curved pattern. So let us look at the second example here. So now what you can see here is when I look at the price of the car versus age of a car here, now that you see notice a different pattern in this case.

So here you can see that so what you will notice here is even though the price of a car is decreasing, or it is showing a downward pattern as the car gets older, but you can see here there is a steep curvature, and it is not appearing linear as it was appearing earlier.

So in this case, it appeared to be a linear trend whereas here it is appearing to have curvature. Similarly, when I look at price of a house versus size of a house this graph appears to be a linear upward trend that is as my size increases, the price appears to increase linearly in terms of the size. Whereas here you can see that there is a curvature or I can say that as my size increases, the price increases according to a curved pattern.

So the next question we ask is, look at the scatter plot and check whether it is linear, or whether it is a curve. For this lecture, we are focusing on linear relationships.

(Refer Slide Time: 05:32)

Statistics for Data Science - I
└ Association between numerical variables
 └ Describing association

Describing association: Variation

Are the points tightly clustered along the pattern?

Price vs. Size

Hand-drawn diagram illustrating the curved pattern of data points:

Google Sheets screenshot:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Size	Price										
2	1.322055115	70.60160154	0.127232452	5.692860196								
3	2.162114767	89.90086904	0.3164282658	25.9216073								
4	3.059880597	53.0297895	-0.4997513213	-0.78911881								
5	4.103162584	68.1770885	-0.278000399	3.269890185								
6	5.048434319	34.0934532	-0.361198449	-29.91525515								
7	6.082453466	60.5538125	-0.417178463	-4.35489587								
8	7.008355666	37.5249529	-0.312170134	-27.3837634								
9	8.1727768615	96.20211401	0.4181957338	31.20349568								
10	9.103590221	86.553339	0.32628383899	21.64463304								
11	10.1718962488	76.26290757	0.4093306071	11.35419922								
12	11.379037982	75.74397095	0.0940578144	10.83526259								
13	11.107785845	48.31077302	-0.2019646395	-16.59853444								
14	11.598265338	80.190492	0.2886544472	-4.71179111								
15	11.411035265	74.96275551	0.1014034141	9.954047155								
16	11.589270128	82.8199837	0.279038247	17.91125354								
17	11.041721674	59.6925953	-0.2079102071	-5.21611311								
18	11.130315674	52.06912476	-0.17916207	-12.8366359								
19	11.614287925	63.36563029	0.304696044	-15.43078059								

Housing data

Housing data full

covariance

Explore

The next question which we want to answer is whether points are tightly clustered around the pattern? Now, if you look at this example, again, I have actually about 100 homes here. I have done the same thing, we have about 100 homes, okay. There are 100 homes I have taken the data of 100 homes here. So, you can see that the data we are talking about is 100 homes, and they are actually tightly clustered around each other.

So, if you look at this scatter plot which we are describing here, these are about 100 homes, and you can see that the data is tightly clustered in the sense that each of these 100 homes are between this range of my prices which could be between 2, 20 or 2 lakhs to 60 lakhs depending on what is the unit of the prices.

So, this is what we refer to as tightly clustered along the pattern. So, what do we mean by a pattern? First, we saw whether it was an upward pattern or a downward pattern, then we checked whether it was a curved pattern or whether it was a linear pattern, this is these 2 are linear patterns, whereas these 2 are curved patterns. Now, we are interested in knowing that along this pattern is my data tightly clustered.

This is the first question or along the pattern is my data variable very high. So, this first example gave me an example of a tightly clustered data whereas the second example is giving me an example of more variable data, both the cases I have an increasing trend which appears linear whereas this is more tightly clustered along the pattern, whereas this is more variable along the pattern.



(Refer Slide Time: 07:44)

Statistics for Data Science -I
└ Association between numerical variables
 └ Describing association

Describing association: Variation



Tightly clustered



Variable

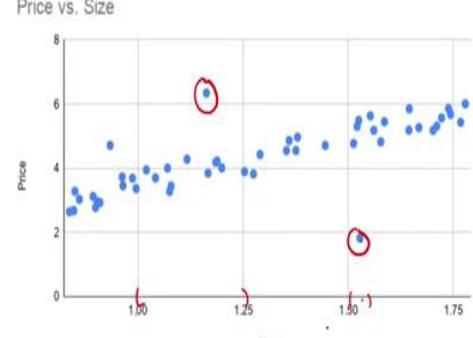


So, that is the next thing which we check this example was tightly clustered. Whereas you can see that this example is more variable or it is not tightly clustered. So, this is the third thing which is check is, how varied is one of the variable or what is the variation exhibited in the variable? This is the next important thing we checked.

(Refer Slide Time: 08:13)

Statistics for Data Science -I
└ Association between numerical variables
 └ Describing association

Describing association: Outliers



Did you find something unexpected?



The last thing which we check is what we refer to as the presence of an outlier. Again refer to the housing data. Now, look at these 3 points. So, there is this in fact, look at these 2 points which I am circling in red and there is this point, this is not an outlier, but now for now, let me just focus

on these 2 points, you can see that all the other points are actually behaving according to a particular pattern whereas these 2 points are away from the regular pattern does exhibited by the other points.

Now, if you look at this point, this tells us about a house which is between 1000 square feet and say 1250 square feet which is actually priced higher than the usual houses in this interval. And what this says is a large sized house which is priced lower than the smallest house also. So, these 2 points are referred to as outliers. So, outlier means or it is not following the pattern which other points exhibit.

(Refer Slide Time: 09:52)

Statistics for Data Science - I
└ Association between numerical variables
 └ Describing association

Section summary



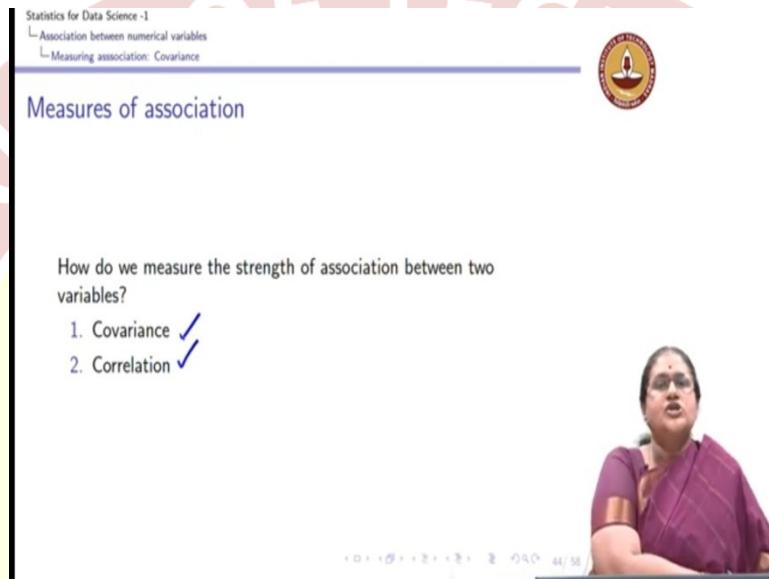
- Describing association
1. Direction
 2. Curvature
 3. Variation
 4. Outliers.
- 

So, when we are looking at association, in summary, the key things which we are trying to look at when we want to talk about association is. First we understand the direction whether it is an upward direction or a downward direction, whether the plot has an upward or a downward direction? Whether it is linear or it is curved? Whether it exhibits variation whether it is tight or whether it is varied?

And finally, whether there are presence of outliers? So, by this time we should know how to plot a scatter plot and look for the association between the variables through visual inspection when we are doing visual inspection these are the 4 key things which we need to take into account.

Statistics for Data Science 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture 4.6
Association between Two Numerical Variables: Covariance

(Refer Slide Time: 00:15)



Statistics for Data Science -I
└ Association between numerical variables
 └ Measuring association: Covariance

Measures of association

How do we measure the strength of association between two variables?

1. Covariance ✓
2. Correlation ✓

So, we have seen how to actually describe association using a scatter plot and we will get an idea between whether an association exists between the two numerical variables under consideration or not. So the next question which we want to ask is; can I quantify this association? In other words, since we are discussing and we are describing numerical variables and we have two numerical variables with us.

The natural question to ask can I come up with a measure of association between the two variables. The two popular measures of association which is used to describe the association between two numerical variables; the first measure is what we refer to as a covariance and the second measure is what we refer to as correlation. So how do we measure the strength of the association?

(Refer Slide Time: 01:21)

Statistics for Data Science -I

- └ Association between numerical variables
- └ Measuring association: Covariance

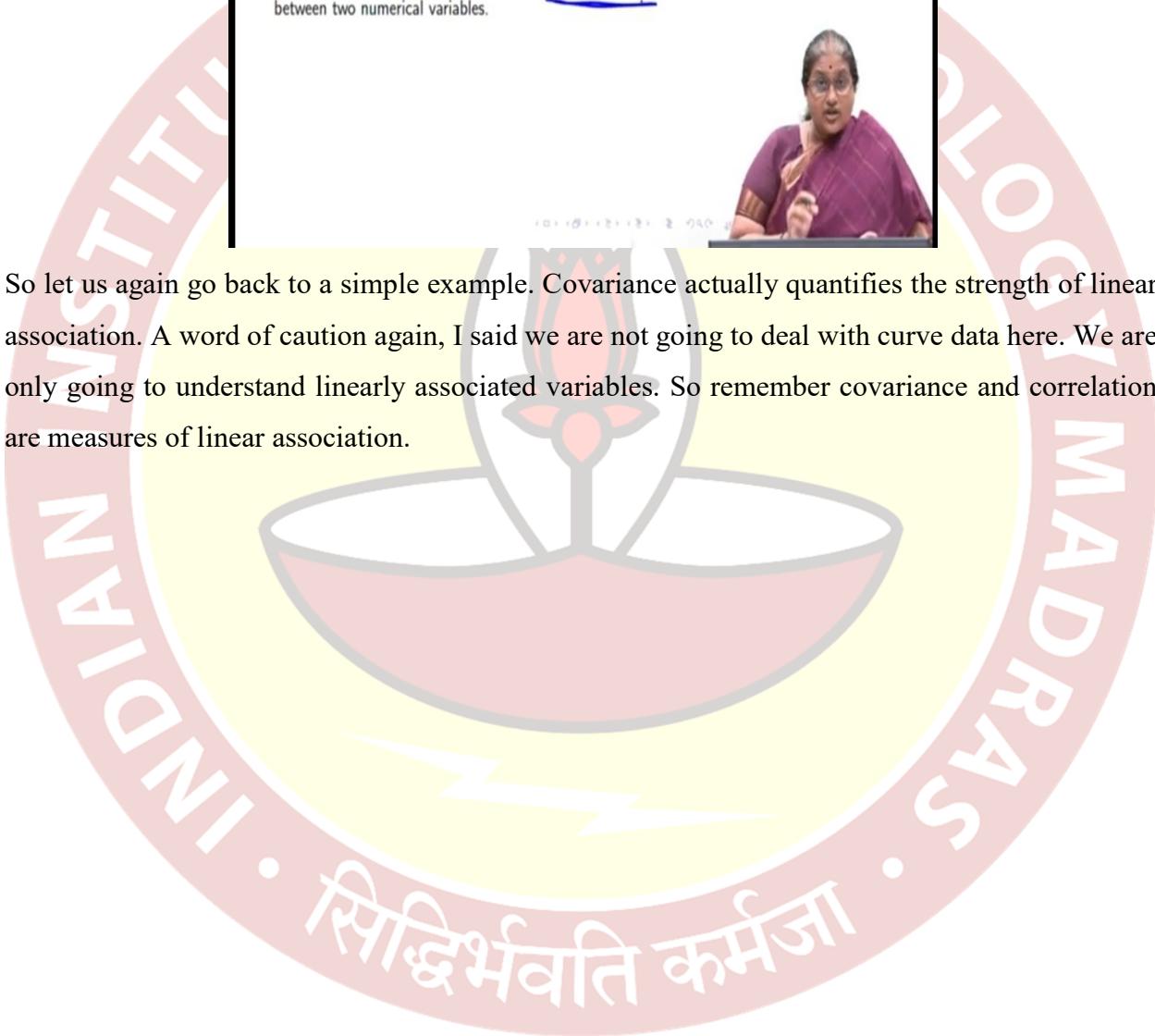
Covariance

Covariance quantifies the strength of the linear association between two numerical variables.



Navigation icons: back, forward, search, etc.

So let us again go back to a simple example. Covariance actually quantifies the strength of linear association. A word of caution again, I said we are not going to deal with curve data here. We are only going to understand linearly associated variables. So remember covariance and correlation are measures of linear association.



(Refer Slide Time: 01:50)

Statistics for Data Science -1
 └ Association between numerical variables
 └ Measuring association: Covariance

Covariance: Example 1

Recall, the association between age and height of a person.

Age (years) x	Height (cms) y	Deviation of x ($x_i - \bar{x}$)	Deviation of y ($y_i - \bar{y}$)
1	75	-2 < 0	-17.6 < 0
2	85	-1 < 0	-7.6 < 0
3	94	0	1.4
4	101	1 > 0	8.4 > 0
5	108	2 > 0	15.4 > 0
$\bar{x} = 3$	$\bar{y} = 92.6$		

A woman in a purple sari is speaking in a video window.

Statistics for Data Science -1
 └ Association between numerical variables
 └ Measuring association: Covariance

Covariance: Example 1

Age (years) x	Height (cms) y	Deviation of x ($x_i - \bar{x}$)	Deviation of y ($y_i - \bar{y}$)
1	75	-2	-17.6
2	85	-1	-7.6
3	94	0	1.4
4	101	1	8.4
5	108	2	15.4

A woman in a purple sari is speaking in a video window.

So let us look at a simple example. Again recall the association between age and height of a person. This is what we have already seen. So now what do we mean by covariance and correlation? So the key observation which we have is the following. So first thing we notice is we want to know, I have the variable here x which is age 1, 2, 3, 4, 5 and I also have the height so I break this. I start with 70, I have 70, 80, 90, 100, 110.

So this is my first point is 70, sorry, this is my first point is 75. Then I have a 85, then I have a 94, then I have a 101 and then I have a 108. This is my scatter plot. Now look at the mean of the first variable. The mean of the first variable is 3 which is nothing but the average of these 5 ages,

which is $1 + 2 + 3 + 4 + 5$ divided by 5. So this is the mean of the first variable. So we know this point, so let me cross out all the points.

So if you look at the mean of the first variable, you can see that in the first variable case there are two points which lie to the left of the mean and two points which lie to the right of the mean and there is one point which lies on the mean. Similarly let us look at the mean of the second variable which is 92.6, so I have a 92.6 here. So let us use a different colour.

So I have this here and I have this here. So what this means is this point lies this point is only mean of the first variable but above the mean of the second variable. Now, these two points are below the mean of both the variables and these two points are above the mean of both the variables. So now you can see that I could have data where I could have had points here.

Now, what are these points? These points would be below the mean of the first variable and above the mean of the second variable. Now, these points would be above the mean of the first variable and below the mean of the second variable. So these so in general what is the idea we are trying to say is if in general I have a scatter plot of x and y and this is the scatter plot which I have.

Suppose I have 100 observations and I have scatter plot of this kind I have the same means, assume I have the same mean. So 3 is my mean here and 92.6 is my mean here. This is my \bar{y} and this is my \bar{x} . So you can see from the scatterplot, I can divide this entire points into 4 regions and let us give different colours to each region. This, the scatter or the points in the orange region are points which are below the mean of both the variables.

The points which are in the yellow region that is these points are points, which are above the mean for both the variables and the points in pink are above the mean for x variable and below the mean for y variable and the points in blue are above the mean for y variable and below the mean of the x variable. Now, why do we care about this? So if I look at the deviation of the orange points from the mean, what do I mean by deviation?

So you can see that all the orange points are lesser than the mean of both x and y variable. So if I look at the orange points my $x_i - \bar{x}$, which is the difference between the point and the mean and $y_i - \bar{y}$ which is the difference between the point on the y axis and its mean, you can see that

since x_i and y_i are lying below \bar{x} and \bar{y} for orange points, both of them are going to be less than 0.

Similarly for the pink points, I will have $x_i - \bar{x} > 0$ and $y_i - \bar{y} < 0$ because y points are lesser than \bar{y} . Here I have both $x_i - \bar{x} > 0$ and $y_i - \bar{y} > 0$. Here I have my $x_i - \bar{x} < 0$ and $y_i - \bar{y} > 0$.

So what is this $x_i - \bar{x}$, $x_i - \bar{x}$ is nothing but the deviation of observation from its mean. Recall, I have two variables. So x is the first variable, y is the second variable, \bar{x} is equal to 3, \bar{y} is equal to 92.6. So my $x_i - \bar{x}$ is 1 - 3, which is -2, 2 - 3 which is -1, 3 - 3 which is 0, 4 - 3 which is 1, 5 - 3 which is 2 the deviation of x_i from its mean.

Similarly, if I look at the deviation of, I can look at 75 - 92.6 which is again -17.6 and similarly I have -7.6 I will have 1.4 and then I have 8.4 and I have 15.4. So you can see that this is the deviation of the points from their respective means. Now why is this again of any interest to us? Now the key point so I have written the deviation of the points from their respective means.

Now what is the interest in this? So here you can notice that the deviation of the first observation on both the variables have the same sign. Second observation again has the same sign, the fourth has the same signed by same sign, I want to say that both of them are greater than 0. Here both of them are less than 0.

(Refer Slide Time: 10:02)

Statistics for Data Science -I
└ Association between numerical variables
└ Measuring association: Covariance

Covariance: Example 2

Variables: Age of a car and price of a car



Age (years) x	Price (INR lakhs) y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$
1	6	-2 < 0	2 > 0
2	5	-1 < 0	1 > 0
3	4	0	0
4	3	1 > 0	-1 < 0
5	2	2 > 0	-2 < 0
3	4		

Now let us look at another example and do the same exercise. So let us look at the deviation here. You can see that again I have 1 - 3 which is - 2; 2 - 3 which is - 1, 0, 1, 2, age; the deviation does not change, 6 - 4 is a 2, 5 - 4 is a 1, 4 - 4 is a 0, 3 - 4 is a -1, 2 - 4 is - 2 and if you look at the deviation of x and y in this example, you can see that this is greater than zero and this is greater than zero, this is greater than zero. When this is greater than zero, I have this less than zero, this less than zero.

So these are the observations which I can make from these 2 examples. So that leads us to a key question that what is the key question we are asking.

(Refer Slide Time: 11:09)

Statistics for Data Science - I
└ Association between numerical variables
└ Measuring association: Covariance

Key observation

- When large (small) values of x tend to be associated with large (small) values of y - the signs of the deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also tend to be same.
- When large (small) values of x tend to be associated with small (large) values of y - the signs of the deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also tend to be different.

So you can look at the key observations that if I am plotting x versus y , if large values of x are associated with large values of y , similarly if small values of x are associated with small values of y , the signs of the deviation, so what do we mean by signs of the deviation?

(Refer Slide Time: 11:38)

Statistics for Data Science - I
└ Association between numerical variables
└ Measuring association: Covariance

Covariance: Example 1

Recall, the association between age and height of a person:

Age (years)	Height (cms)	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$
1	75	-2 < 0	-17.6 < 0
2	85	-1 < 0	-1.6 < 0
3	94	0	14
4	101	1 > 0	84.7 > 0
5	108	2 > 0	154.7 > 0
$\bar{x} = 3$	$\bar{y} = 92.6$		

If you look at this case the first example you saw the case. In the first example, we saw that large values of age. What are the large values of age? 4 and 5 large values of height 101, 108; 3, 4, 5 can be looked as large values of age, and they are associated with large values of height. Small

values of age are associated with small values of height. Hence, you can see that the deviation of both the variables have the same size. Here, the deviation is less than 0. Here also, the deviation less than 0. Here again it less than 0. It is less than 0. This is greater than 0, greater than 0, greater than 0, greater than 0.

(Refer Slide Time: 12:33)

Statistics for Data Science - I
└ Association between numerical variables
 └ Measuring association: Covariance

Covariance: Example 2

Variables: Age of a car and price of a car

Age (years) x	Price (INR lakhs) y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$
1	6	-2 < 0	3 > 0
2	5	-1 < 0	1 > 0
3	4	0	0
4	3	1 > 0	-1 < 0
5	2	2 > 0	-2 < 0
3	4	0	0

Whereas in the next example, you can see that the small value is associated with the large value of price and the large value is associated with small value. Similarly 2 is associated with the next larger value and 4 is associated with a small value and as a consequence, you can see that the deviation of x is negative but the deviation of y is positive. Similarly when the deviation of x is positive, the deviation of y is negative. In other words the deviation of the variables are of different signs. So, how do we use this observation?

(Refer Slide Time: 13:20)

Statistics for Data Science -1
└ Association between numerical variables
└ Measuring association: Covariance

Key observation

When large (small) values of x tend to be associated with large (small) values of y - the signs of the deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also tend to be same.

When large (small) values of x tend to be associated with small (large) values of y - the signs of the deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also tend to be different.

x y Deviation (x, y)

Large (Small) Large (Small) - Same.

Large (Small) Small (Large) - different

Remember we wanted to quantify the association. So when we look at this, so when the large value tends to be associated with small or small tends to be associated with large, we see that the signs and the deviation are different whereas if the large is associated with large and small is associated with small, we see the signs of deviation to be same.

So to summarize if the large, so x large is associated with y large or x small is associated with y small. The deviation of x and y are same, the signs. Similarly if x large is associated with y small and x small is associated with y large, we find the deviation signs to be different. This is one key observation which we have. So the question is now if the deviation signs are the same, then if I take a product of these deviation, it could give me a measure of the covariance or association. So how do we look at that?

(Refer Slide Time: 14:49)

Statistics for Data Science -1
└ Association between numerical variables
 └ Measuring association: Covariance

Covariance: Example 1

(+)(+)
(-)(-)

Age x	Height y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	-2	-17.6	35.2
2	85	-1	-7.6	7.6
3	94	0	1.4	0
4	101	1	8.4	8.4
5	108	2	15.4	30.8

So if I have for example, look at my first example. I saw that the signs of the deviations are same. In other words, I could have if x is a positive deviation, my y is also a positive deviation. If x is a negative deviation my y is also a negative deviation. You can see both are negative in this case, both are positive in this case.

So if I look at the product of the deviations, I am going to have a positive sign. If I look at the product of the deviation because both the deviation $x_i - \bar{x}$ and $y_i - \bar{y}$ are of same size. When I look up the product of these deviations, I will get a positive sign and you can see that $(-2) \times (-17.6)$ is 35.2, $(-1) \times (-7.6)$ is 7.6. Here, I have a 0, I have 1×8.4 is 8.4, 2×15.4 is 30.8.

So I have the product of deviations. All my deviations is positive in this case. There be a chance of one being positive or negative, there could be a chance but for in this simple example, I can see that the product of all my deviation is positive.

(Refer Slide Time: 16:21)

Statistics for Data Science -1
└ Association between numerical variables
 └ Measuring association: Covariance

Covariance: Example 2

$-ve +ve$ $+ve -ve$ $\boxed{-ve}$

Age x	Price y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	6	-2	2	-4
2	5	-1	1	-1
3	4	0	0	0
4	3	1	-1	-1
5	2	2	-2	-4



The slide shows a table of data points for Age (x) and Price (y). The deviations from the mean are calculated: Age (x) has deviations of -2, -1, 0, 1, 2; Price (y) has deviations of 2, 1, 0, -1, -2. The covariance is calculated as the sum of the products of corresponding deviations: (-2*2) + (-1*1) + (0*0) + (1*-1) + (2*-2) = -4 - 1 + 0 - 1 - 4 = -10. Handwritten annotations show the terms $-ve +ve$, $+ve -ve$, and $\boxed{-ve}$ above the table.

Similarly, let us look at the other extreme example. Here my x deviation is negative, this is positive so the product, so I either have a negative x and a positive y deviation or a positive x and a negative y deviation. So here these 2 are negative x with a positive y deviation (whether these two) whereas these 2 are positive x deviations with negative y deviations.

I know the product of these deviations is always going to have a negative sign. So you can see that -2×2 is a -4 . I have a -1 which is -1×1 , 1×-1 is again -1 , 2×-2 is -4 . So I can say that the key observations we are trying to make is when I am coming up with a product of deviations.

(Refer Slide Time: 17:29)

Statistics for Data Science -1
 └─Association between numerical variables
 └─Measuring association: Covariance

Covariance: Example 1

		(+) (+) (-) (-)	(+/-) (-/+)
Age	Height	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$
1	75	-2	-17.6
2	85	-1	-7.6
3	94	0	1.4
4	101	1	8.4
5	108	2	15.4



When I have data of this kind, this is one extreme and the other extreme is I have that data where my deviations are when it is positive, it is positive and when the deviation is negative, it is negative. That is, I have a match of deviation. There could be a case where I could have a negative deviation and a positive deviation. We look at it very soon.

(Refer Slide Time: 17:54)

Statistics for Data Science -1
 └─Association between numerical variables
 └─Measuring association: Covariance

Key observation



- When large (small) values of x tend to be associated with large (small) values of y - the signs of the deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also tend to be same.
- When large (small) values of x tend to be associated with small (large) values of y - the signs of the deviations, $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will also tend to be different.

x y Deviations (x, y)
 Large (small) Large (small) - Same.
 Large (small) Small (large) - different

Covariance: Example 2

-ve +ve
 +ve -ve



Age x	Price y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	6	-2	2	-4
2	5	-1	1	-1
3	4	0	0	0
4	3	1	-1	-1
5	2	2	-2	-4



So now you can see that the key idea we have defined is, when I have this deviation the product of the deviation could be positive or it could be negative. In extreme cases here, I have all the products of the deviations negative.

(Refer Slide Time: 18:15)

Covariance: Example 1



Age x	Height y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	-2	-17.6	35.2
2	85	-1	-7.6	7.6
3	94	-0.5	1.4	-0.7
4	101	1	8.4	8.4
5	108	2	15.4	30.8



And in the earlier example, I had all the products to be positive, but I could have a situation where for example if this was instead of 3, if it was 3.5, sorry, if it was a different age, I could have an example where just a hypothetical situation that this instead of 0 was or say some 0.5, we should make appropriate changes here. In that case I could have so this was or a - 0.5, I could

have here a case where the product of - 0.5, so I will have a 1.4×0.5 which is about 7, but it would have a negative sign.

This could be rare but mostly I would find cases which are positive. I might have a few cases which are negative product deviation. So how would I measure the strength of an association? Remember by strength of an association we are seeking the answer to: do both the variables increase together, do they decrease together is the association linear, do I have outliers? These are the questions which we are trying to answer.

So one way is, if as in the earlier example, if this was a -7 then you can see that this negative would have cancelled out with the positives, if everything were positive, then I could have told a story. So one way to quantify this measure is take the sum of all the deviations, the sum of these deviations will tell us what is the strength of the association.

Now the sum over, so I am just not looking at the sum of the deviations, I need to look at the sum of the deviations which would be my numerator and I have to divide it by the number of observations.

(Refer Slide Time: 20:36)

Covariance

	x_1	y_1
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
\vdots	\vdots	\vdots
$N(n)$	$x_{N(n)}$	$y_{N(n)}$

Definition

Let x_i denote the i^{th} observation of variable x , and y_i denote the i^{th} observation of variable y . Let (x_i, y_i) be the i^{th} paired observation of a population (sample) dataset having $N(n)$ observations. The Covariance between the variables x and y is given by

► **Population covariance:** $\text{Cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{N}$

► **Sample covariance:** $\text{Cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

So we formally define what is a covariance measure. So if I have my dataset where I have my 1, 2, 3, these are the number of my observations. I have a N if I refer to a population this is from our earlier lectures. I would have a small n if I am referring to my dataset which comes from a

sample. I have my x variable. I have my y variable, x_1, x_2, \dots, x_N or x_n are the N observations of the first variable, y_1, y_2, \dots, y_N or y_n are the observations from the second variable.

Then $x_i - \bar{x}$ is the deviation of my first variable from its mean, $y_i - \bar{y}$ is a deviation of my second variable from its mean. This is the product of deviation. I sum up the product of deviations over all possible values and divide by the total number of values I get what I refer to as the population covariance.

Similarly, if I sum up the product of deviations over all the sample values, n is the sample value I divide it by $n - 1$. Recall when we defined the sample standard deviation and the sample variance also we divided it by $n - 1$ and at that point of time, I said when you are referring to a population you divide it by N, if you are referring to a sample you divide it by $n - 1$.

We use the same thing when we refer to a sample covariance. We divide the numerator with $n - 1$ observations and this quantity is referred to as the sample covariance. So how do I compute the sample covariance?

(Refer Slide Time: 22:50)

Statistics for Data Science - 1
└ Association between numerical variables
 └ Measuring association: Covariance

Covariance: Example 1

Age x	Height y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	-2	-17.6	35.2
2	85	-1	-7.6	7.6
3	94	0	1.4	0
4	101	1	8.4	8.4
5	108	2	15.4	30.8
				82

► Population covariance: $\frac{82}{5} = 16.4$ $N=5$ $\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y}) = 82$

► Sample covariance: $\frac{82}{4} = 20.5$ $n=5$

So now you can see that for the first example, I add all my quantities. I get 82. What is this 82? $82 = \sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y})$. That is what is 82 of these 5 observation is equal to 82, which is given by this. So if I am interested in finding the population covariance my N is equal to 5. I divide 82 by 5 and I have 16.4 which will give me my population covariance whereas if I want to

find out my sample covariance, I know again n equal to 5. I divide 82 by n - 1 which is a 4 I get a sample covariance of 20.5.

(Refer Slide Time: 23:57)

Statistics for Data Science -1
└ Association between numerical variables
└ Measuring association: Covariance

Covariance: Example 2

Age <i>x</i>	Price <i>y</i>	Deviation of <i>x</i> $(x_i - \bar{x})$	Deviation of <i>y</i> $(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	6	-2	2	-4
2	5	-1	1	-1
3	4	0	0	0
4	3	1	-1	-1
5	2	2	-2	-4
				$\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y}) = -10$

- Population covariance: $\frac{-10}{5} = -2$
- Sample covariance: $\frac{-10}{4} = -2.5$

So you can see the same thing for the second example, you can see that when I add up my deviations, I have -10 as the sum. So $\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y}) = -10$. The sign is very important. Again, if I am interested in that population covariance, I divide it by 5. If I am interested in the sample covariance, I divide it by 4, you can again see there is a sign the population covariance in the second example is -2 where a sample covariance is -2.5.

So what is the sample covariance give? It gives us a quantified measure when two variables are moving in the same direction, the covariance is a positive measure whereas when two variables are moving in opposite direction, what do we mean by opposite direction? As one variable increases the other variable decreases then my sample covariance is negative. So this is how you quantify the covariance. How does Google Sheets give you the covariance?

(Refer Slide Time: 25:18)

The image shows two screenshots of a Google Sheets document titled "Association between numerical variables".

Top Screenshot: A table with columns A through F. Rows 1-6 contain data: Age (1, 2, 3, 4, 5), Height (75, 85, 94, 101, 108), xdev (-2, -1, 0, 1, 2), ydev (-17.6, -7.6, 1.4, 8.4, 15.4), and ProdDev (35.2, 7.6, 0, 8.4, 30.8). Row 7 has values 3, 92.6, and 82. Row 8 is empty. Row 9 has values 1 and 6.

	A	B	C	D	E	F
1	Age	Height	xdev	ydev	ProdDev	
2	1	75	-2	-17.6	35.2	
3	2	85	-1	-7.6	7.6	
4	3	94	0	1.4	0	
5	4	101	1	8.4	8.4	
6	5	108	2	15.4	30.8	
7	3	92.6			82	
8						
9	1	6				

Bottom Screenshot: A table with columns C through H. Rows 1-6 contain data: xdev (-2, -1, 0, 1, 2), ydev (-17.6, -7.6, 1.4, 8.4, 15.4), and ProdDev (35.2, 7.6, 0, 8.4, 30.8). Row 7 has value 82.

	C	D	E	F	G	H
1	xdev	ydev	ProdDev			
2	-2	-17.6	35.2	16.4	16.4	
3	-1	-7.6	7.6	20.5	20.5	
4	0	1.4	0			
5	1	8.4	8.4			
6	2	15.4	30.8			
7			82			

Let us go back to Google Sheets. Go to this sheet here. So let us look at the first example here. So I have my first example, which is given by this quantity here. So you can you recall. This was again my age in years and this was the price, sorry, this was the height in centimetres. This is 3 is my \bar{x} , so I can find my deviation of x.

So I will call that x deviation or x dev. That is what I am going to refer to and I will also refer to as y deviation, $y_i - \bar{y}$ as y dev. So x deviation, I can find out what is my x deviation, x deviation is going to be x_i which is my age - my mean.

Since I am going to have the same mean I am just going to, this is what I have and you can see that I have $x_i - \bar{x}$ and $y_i - \bar{y}$. These are my deviation. This is precisely what we had in our, you can see this is for the first example I have -2, -17.6. So that is what my Google Sheets give me. I am doing the example right from beginning. I can find the product of the deviation and that is nothing but I can find the product of the deviation.

The product of the deviation, I can write that down and you can see that this product of the deviation is precisely what we have here, 35.2 up to 30.8. Now I can find out what is the sum of these deviations and if I divide the sum by I can divide this sum by 5, I can divide this by 5 I get my population covariance. I divide this sum by 4, I get my sample covariance.

Now, there is a function n in our Google Sheets. If you go to a Google sheet and type covariance, you can see that there are 3 functions available which is covariance of a dataset, covariance.P, which is covar and covariance P, you make a population covariance of the datasets whereas covariance.S gives me the sample variance of a dataset. So let us see what covariance.P gives me. So I have covariance.P. So to find out the covariance, my first variable, I choose the first variable here.

The second variable is height. I choose that and then you can see that the covariance or covariance.P of my variables gives me precisely the population covariance which we have worked out from first principles. Similarly, the covariance sample covariance of both the dataset so again, I choose my age and I choose the height.

The sample covariance is equivalent to dividing this 82 by $n - 1$ which is 4, gives me the sample covariance. I can do the same thing for my next example also. We will just quickly go about it because we have already done it.

(Refer Slide Time: 29:39)

The image displays two screenshots of a Google Sheets document titled "Association between numerical variables".

Screenshot 1: This screenshot shows a portion of a covariance matrix. The columns are labeled A, B, C, D, E, F and the rows are numbered 6 through 14. The matrix values are as follows:

	A	B	C	D	E	F
6	5	108	2	15.4	30.8	
7	3	92.6			82	
8	age	price				
9	1	6	-2	2		
10	2	5	-1	1		
11	3	4	0	0		
12	4	3	1	-1		
13	5	2	2	-2		
14	3	4				

Screenshot 2: This screenshot shows another portion of the covariance matrix. The columns are labeled C, D, E, F, G, H and the rows are numbered 7 through 14. The matrix values are as follows:

	C	D	E	F	G	H
7			82			
8						
9	-2	2	-4	-2	-2.5	
10	-1	1	-1			
11	0	0	0			
12	1	-1	-1			
13	2	-2	-4			
14			-10			
15						



Covariance: Example 2

Age x	Price y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	6	-2	2	-4
2	5	-1	1	-1
3	4	0	0	0
4	3	1	-1	-1
5	2	2	-2	-4
				-10

- ▶ Population covariance: $\frac{-10}{5} = -2$
- ▶ Sample covariance: $\frac{-10}{4} = -2.5$

$$\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y}) = -10$$



This is again this is my age of a car and this is the price of a car. So I can find out what is my, I will just do the deviations quickly. So this is going to be $x_i - \bar{x}$. I can freeze this. So we notice here again the deviations of x and y are of different size. I take the product of these 2 and I sum it up. I have a -10. Again this is very, this is what we did manually and you can see that, that is equal to -10 here.

So now what is my population covariance? My dataset is going to be age and price which you can see is -2 and -2 = $-10/5$ where 5 is the population size and similarly my sample covariance is again going to be again, if I compute my sample covariance for the same pairs of variable, I get it - 2.5 which is $-10/4$ and you can see that my sample covariance is - 2.5 which is consistent and the same would be obtained when we did it manually.

(Refer Slide Time: 31:29)

Statistics for Data Science -1
↳ Association between numerical variables
↳ Measuring association: Covariance

Units of Covariance

	PC	SC	
First dataset	16.4	20.5	
Second dataset	-2	-2.5	
your Age	years	years	
Height	cm	cm	
Cars			

► The size of the covariance, however, is difficult to interpret because the covariance has units.

Statistics for Data Science -1
↳ Association between numerical variables
↳ Measuring association: Covariance

Covariance: Example 1

Age x	Height y	Deviation of x $(x_i - \bar{x})$	Deviation of y $(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	-2	-17.6	35.2
2	85	-1	-7.6	7.6
3	94	0	1.4	0
4	101	1	8.4	8.4
5	108	2	15.4	30.8
				82

► Population covariance: $\frac{82}{5} = 16.4$ $N=5$ $\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y})$

► Sample covariance: $\frac{82}{4} = 20.5$ $N=5$

So now the question is what the covariance measure gives us is the size of an association. So in earlier case I had two sizes. So when we are looking at population or so for now, let us restrict ourselves to sample covariance. So for my second dataset and my first dataset, my population covariance and my sample covariance. So if you go back, my population covariance in this first dataset was 16.4 and 20.5 whereas here this was a - 2 and - 2.5.

So you can see that this is a positive measure whereas here this is a negative measure. Now the question is when you look at the first dataset my variables here was again age and height. My x variable was age, this was measured in years, my height y, it was measured in centimetres. If x is

measured in age, I know \bar{x} is also takes the same units as my original variable so \bar{x} is also in years.

If \bar{x} is in years, $x_i - \bar{x}$ is also in years. Similarly if each y_i is in centimetres, \bar{y} is also in centimetres so $y_i - \bar{y}$ is in centimetres. So this $x_i - \bar{x}$, this deviation is in years and $y_i - \bar{y}$ is in centimetres. Hence, the product of these two variables is actually going to, the units are going to be a product of the units. There is a product of the units actually is a difficult thing to articulate.

Similarly when I look at this second dataset, my x_i is again in years. So my $x_i - \bar{x}$ is in years whereas my $y_i - \bar{y}$ for my second dataset is in currency or is it in lakhs of rupees, Indian national rupees in lakhs. So again, the product is going to have a unit of measure. Each product has a unit of measure. The summation over all the observations would also take the same units of measure.

So because of this it actually is very difficult to interpret the covariance measure. So the next natural thing to ask is can I have a unit-less measure so that it becomes easier for me to interpret the strength of association between two variables. The answer is yes.

(Refer Slide Time: 34:49)

Statistics for Data Science - 1
└ Association between numerical variables
 └ Measuring association: Covariance

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

Units of Covariance

- ▶ The size of the covariance, however, is difficult to interpret because the covariance has units.
- ▶ The units of the covariance are those of the x-variable times those of the y-variable.

57 / 58

So the question is, is there a unit of measure since the units of covariance are those of the variable of x times the variable of y and it becomes difficult to interpret. Thought it gives a good measure of strength of association, very natural strength of association, the question is do I have

another measure which can help in interpreting the strength of association better. Answer is yes and that is the measure which we refer to as a correlation measure which we will be seeing next.

(Refer Slide Time: 35:26)

Statistics for Data Science -1
└ Association between numerical variables
 └ Measuring association: Covariance

Section summary

+ve
-ve

1. Introduced the measure of covariance
2. How to interpret the covariance measure



Navigation icons: back, forward, search, etc.

So in summary what we have learned so far is we introduced the measure of covariance. We saw what was the inclusion behind coming up with this measure of covariance and we saw how to interpret this covariance measure in a sense if it is positive, then in a sense we say that two variables might be moving up in the same direction and we assume linear because both covariance and correlation are measures of linear association.

And if the covariance is negative, it means that if 1 variable is moving in an up direction, the other variable is moving in the down direction, but however interpretation is difficult, hence, we seek another measure and that measured is what we refer to as the correlation measure.

Statistics for Data Science - 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Management, Madras
Lecture 4.7

Association between two numerical variables – Correlation

(Refer Slide Time: 0:24)

Statistics for Data Science -1
└ Association between numerical variables
└ Measuring association: Covariance

INSTITUTE OF MANAGEMENT TECHNOLOGY
MADRAS

Units of Covariance

- ▶ The size of the covariance, however, is difficult to interpret because the covariance has units.
- ▶ The units of the covariance are those of the x-variable times those of the y-variable.

57 / 77

So, we understand now another measure of association again when I say association, I mean linear association between two numerical variables. You already seen the measure of covariance, but recall when we talked about covariance, we said that the covariance is difficult to interpret, because the covariance has units.

(Refer Slide Time: 0:41)

Statistics for Data Science -1
└ Association between numerical variables
└ Measuring association: Correlation

INSTITUTE OF MANAGEMENT TECHNOLOGY
MADRAS

Correlation

- ▶ A more easily interpreted measure of linear association between two numerical variables is correlation.
- ▶ It is derived from covariance.
- ▶ To find the correlation between two numerical variables x and y divide the covariance between x and y by the product of the standard deviations of x and y . The Pearson correlation coefficient, r , between x and y is given by

$$r = \frac{\text{Cov}(x, y)}{S_x S_y}$$

60 / 77



Correlation

- A more easily interpreted measure of linear association between two numerical variables is **correlation**
- It is derived from covariance.
- To find the correlation between two numerical variables x and y divide the covariance between x and y by the product of the standard deviations of x and y . The Pearson correlation coefficient, r , between x and y is given by

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\text{cov}(x, y)}{s_x s_y}$$

60 / 77

The screenshot shows a Google Sheets document with the title "Association between numerical variables". The table has four columns: A (empty), B ("Size (1000 Square feet)", bolded), C ("Price (INR Lakhs)", bolded), and D (empty). The data consists of 11 rows of house size and price pairs. The last row is a summary row with values 10 and 2.5 under columns B and C respectively. The formula bar at the bottom shows the range "Housing data!A2:D11".

A	B	C	D
1	Size (1000 Square feet)	Price (INR Lakhs)	
2	1	0.8	68
3	2	1	81
4	3	1.1	72
5	4	1.3	91
6	5	1.6	87
7		1.8	56
8	7	2.3	83
9	8	2.3	112
10	9	2.5	93
11	10	2.5	98

So, we are going to another measure of association and this is what I term, the correlation between two numerical variables. What is a correlation? It is a, again it is another measure of linear association between two numerical variables, it is derived from the concept of covariance, we have already introduced the concept of covariance. Now, how do I find the correlation between two numerical variables, let me call those two numerical variables x and y .

Again going back to our example, you can see that I have this example here, where I have size and price of a house, there are 15 observations. So, if you can look at this, let me zoom it a bit more. So, you can see that I have the size and price. So, x is my explanatory variable, which is the size of the house, y is my response variable which is the price of the house.

So, if I want to know the correlation between these two variables, I can represent it by r or some books represented by ρ . The Pearson correlation coefficient r between variables x and y , r_{xy} I can drop this x and y also. And just represent it by r , the Pearson correlation coefficient is given by r denote the covariance between x and y divided by the product of the standard deviations of x and y i.e. $r = \frac{\text{Cov}(x,y)}{s_x s_y}$

Since, I have already said that it is derived from the covariance, a more formal way of defining the correlation is the following, $r = \frac{\sum_{i=0}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=0}^n (y_i - \bar{y})^2}}$. This is my n here, I am summing it up in the case of my example, n equal to 15. Because I have the data on 15 homes I am looking at the sizes and prices of these 15 homes I divided by the square sum of deviations, this is the sum of the, this is the square deviation sum of square deviations of x and this is the sum of square deviations of y.

In other words, I also have that this is equivalently I can write that $r = \frac{\text{cov}(x,y)}{s_x s_y}$. So, this is how we compute our correlation between two numerical variables. Now, why is it important? Now, when you look at this term $\text{cov}(x, y)$.

(Refer Slide Time: 4:21)

NDIA

Statistics for Data Science -1
└ Association between numerical variables
└ Measuring association: Covariance

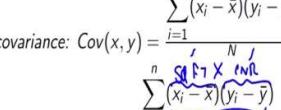


Covariance

Definition

Let x_i denote the i^{th} observation of variable x , and y_i denote the i^{th} observation of variable y . Let (x_i, y_i) be the i^{th} paired observation of a population (sample) dataset having $N(n)$ observations. The Covariance between the variables x and y is given by

$$\text{Population covariance: } \text{Cov}(x, y) = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{n}$$



$$\text{Sample covariance: } \text{Cov}(x, y) = \frac{\sum_{i=1}^{n-1} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

I know that the covariance term, again go back to the where we defined a covariance, we saw that when we looked at the covariance term, this is how we defined a covariance term, we looked at the deviation. Now, the deviation for example, if x is measured in terms of square feet, the deviation is also going to have the unit of square feet, y is measured in terms of INR.

The deviation also is in terms of the currency which is in lakhs of rupees. So, this correlation covariance term has a unit which is the product of the unit of this variable and this variable which is square feet into Indian National Rupees.

(Refer Slide Time: 5:14)

Statistics for Data Science - I
↳ Association between numerical variables
↳ Measuring association: Correlation

Correlation

- ▶ A more easily interpreted measure of **linear association** between two numerical variables is **correlation**
- ▶ It is derived from covariance.
- ▶ To find the correlation between two numerical variables x and y divide the covariance between x and y by the product of the standard deviations of x and y . The Pearson correlation coefficient, r , between x and y is given by

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\text{cov}(x, y)}{s_x s_y}$$

*size price
sq ft x m²*

60 / 77

Now, when I look at the correlation for the same variable, I have defined the correlation as the following. Covariance, now, what are the units of covariance, covariance in the example, the units of this covariance is square feet, which is the unit of size which is my variable x , and then I have which is INR in lakhs of rupees, which is the unit of price, which is my response variable y .

Now, if you look at the standard deviation of x , standard deviation of x is also going to have the same units as that of x , standard deviation of y is going to have the same units as that of y , which is in Indian National Rupees. So, you can see that the units cancel off when I talk about a correlation measure.

(Refer Slide Time: 6:09)



Remark

The units of the standard deviations cancel out the units of covariance

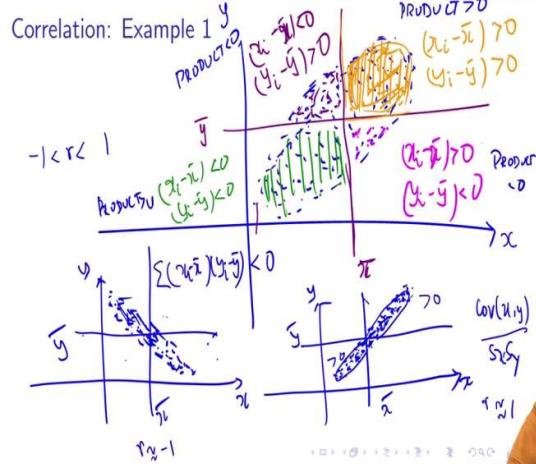
Remark

It can be shown that the correlation measure always lies between -1 and +1

$$-1 \leq r \leq +1$$

So, this measure of correlation is a unit less quantity. But, we need to also remember that it is a measure of linear association, it can be shown that this correlation coefficient always lies between +1 and -1. So now, we have a covariance measure and a correlation coefficient.

(Refer Slide Time: 6:46)



Now, how do we use this correlation coefficient to tell about the strength of the association between my variables? So, recall when we looked at the covariance matrix, we started by what we said was a scatter plot matrix. In the scatter plot matrix, I have my explanatory variable, which is on the x axis, I have my response variable which is on the y axis.

So, if I have a scatter of this kind between these two variables, I am just looking at one of the quadrant, but I could have in both the quadrant but I am just for exposition sake, I am just looking at this quadrant. And suppose I have my \bar{x} , so x varies from this point to this point. So, suppose this is my \bar{x} that is the mean of my first variable.

And suppose this is my \bar{y} , suppose this is my \bar{x} , and this is my \bar{y} , we further said that we could split this entire scatter into four pieces. In this piece, if I am going to have all my points orange. This is this piece, that is where my, so if I am going to share this region or this points, so all the points in this region are basically points where my $x_i - \bar{x}$ and my $y_i - \bar{y}$ are both greater than 0 because both all the points are greater both my x point and my y point are greater than the respective means.

Now, in this green area, green shaded area, my $x_i - \bar{x}$ and my $y_i - \bar{y}$ are both less than 0 because both the x_i points and the y_i points are less than their respective means. Now, if I look at the purple area, which is this scatterplot, I find the y_i 's are greater than the \bar{y} , whereas my x_i is less than the \bar{x} . And in this final area, which is say, purple the smaller area here, I have here $x_i - \bar{x} > 0$, whereas $y_i - \bar{y} < 0$.

Now, let us understand why this correlation metric or this correlation coefficient becomes very important. Now, suppose I have a dataset of this kind where I have a tightly clustered data. So, remember when we talked about association, we wanted to see the direction. And whether it is tightly clustered, whether there is a variability all these points, suppose it is a tightly clustered data, for the same data point, if I continue with the exercise I have done before, this is my \bar{x} , this is my \bar{y} .

So, you can see that, in this I have very little points which are here, but majority of the points lie in this area, and in this area, the product. So, when I look at the product of the deviation, it is greater than 0 in this quadrant, the product of the deviation is again greater than 0 in this quadrant, hence the sum of the product of deviation, which is $\sum(x_i - \bar{x})(y_i - \bar{y})$, because a product is always greater than 0, the sum is going to be greater than 0. Hence, this type of a pattern will always result in a positive covariance.

Whereas if I have my data, which is of this kind now, I have my x here, I have my y here, and my data is of this kind. Suppose, this is my scatterplot, again this is my \bar{x} , this is my \bar{y} , this is my x , this is my y , this is my \bar{x} , I have the same \bar{y} , now if you notice, this scatterplot, you can see that the product here I have a product.

So, if I look at the product of these coefficient, the product of the deviations is greater than 0, here the product is going to be greater than 0 in this quadrant, here the product is going to be less than 0. Here, also the product is going to be less than 0. So, what we notice here is the product of the deviation.

So, $(x_i - \bar{x})(y_i - \bar{y})$ in this portion is going to be less than 0, in this portion is also less than 0, I do not have any points in these two areas. So, $\sum(x_i - \bar{x})(y_i - \bar{y}) < 0$. And this resulted in a negative covariance measure.

Now, if I am going to divide them by their respective standard deviations, I get my correlation measure. So, the correlation measure always is between -1 and +1. How do I interpret this correlation measure? If there is a, so, if the data is of this kind, my covariance I know is going to be positive I divided by the standard deviations, and I see that if my data is of this kind, then my correlation measure is very close to 1. Whereas, if my data is this kind, my correlation measure is very close to -1, so this is a perfect linear relationship between x and y in the positive direction. This is a perfect linear relationship between x and y in the negative direction.

(Refer Slide Time: 14:02)

Statistics for Data Science -1
I – Association between numerical variables
L – Measuring association: Correlation

Correlation: Example 1

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{82}{\sqrt{10} \sqrt{677.2}}$$

Age <i>x</i>	Height <i>y</i>	sq.Devn of <i>x</i> $(x_i - \bar{x})^2$	sq.Devn of <i>y</i> $(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	75	-2 ² = 4	17.6 ² = 309.76	35.2
2	85	-1 ² = 1	57.76	7.6
3	94	0 ² = 0	1.96	0
4	101	1 ² = 1	70.56	8.4
5	108	2 ² = 4	237.16	30.8
$\bar{x} = 3$	$\bar{y} = 92.6$	$\sum(x_i - \bar{x})^2 = 10$	$\sum(y_i - \bar{y})^2 = 677.2$	$\sum(x_i - \bar{x})(y_i - \bar{y}) = 82$

► $s_x = 1.58$, $s_y = 13.01$

► $r = \frac{82}{\sqrt{10} \sqrt{677.2}}$ OR $\frac{20.5}{1.58 \times 13.01} = 0.9964$

$\frac{Cov(x,y)}{s_x \times s_y} = \frac{20.5}{1.58 \times 13.01}$

Navigation icons: back, forward, search, etc.

Before I go to this dataset, let us work on a small hypothetical example. Again, I had x, which was my age, y, which was the height I wanted to explain whether as people grow older, their heights increase or decrease, I wanted to know what is the association, so we found out what was the deviation, so I knew the mean here was 3, we computed the mean here, this is what we already did. This is what we did in our, so my mean was a 3, and the mean here was a

92.6, so the deviation was 1 - 3, which was -2, the deviation was 2 - 3 a - 1, 4 equal to -2^2 , -1^2 is 1, so this is the sum or the square deviation is what is given here.

Similarly 0, 1, and 2, this is giving me the square deviation, similarly 75 - 92.6; I also computed that as -17.6×-17.6^2 is equal to 309.76. And this is giving me the square deviation of each of my y observation. This is the sum of square deviations. This is the sum of square deviations of y. This is the sum of the product of the deviations, so if I apply my first formula, the correlation coefficient $r = \frac{\sum_{i=0}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=0}^n (y_i - \bar{y})^2}}$. Sum of total sum is equal to 82, divided by $\sqrt{10} \times \sqrt{677.2}$.

Using the second formula, $r = \frac{Cov(x,y)}{s_x s_y}$. What is the covariance of this value, we have already computed the covariance, we know the covariance is 20.5 covariance between x and y, this is something which you have already computed, I compute what is the s_x , which is 1.58, s_y , which is 13.01. I divide by 1.58×13.01 and get 0.9964. Hence, this 0.9964, which is very very close to 1 captures the strength of the linear relationship between age and height. So, the strength of the linear relationship, this is a positive very strong linear relationship between age and height.

(Refer Slide Time: 17:23)

Statistics for Data Science - I
 L- Association between numerical variables
 L- Measuring association: Correlation

Correlation: Example 2

Years WR ✓

Age	Price	sq. Devn of x $(x_i - \bar{x})^2$	sq. Devn of y $(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	6	$-2^2 = 4$	$2^2 = 4$	-4
2	5	$-1^2 = 1$	$1^2 = 1$	-1
3	4	$0^2 = 0$	$0^2 = 0$	0
4	3	$1^2 = 1$	$-1^2 = 1$	-1
5	2	$2^2 = 4$	$-2^2 = 4$	-4
$\bar{x} = 3$	$\bar{y} = 4$	10	10	-10

$\rightarrow s_x = 1.58, s_y = 1.58$
 $\rightarrow r = \frac{-10}{\sqrt{10} \times \sqrt{10}} \text{ OR } \frac{-2.5}{1.58 \times 1.58} = -1$

$Cov(x, y) = -2.5 = -10/4$
 $s_x \times s_y = 1.58 \times 1.58$

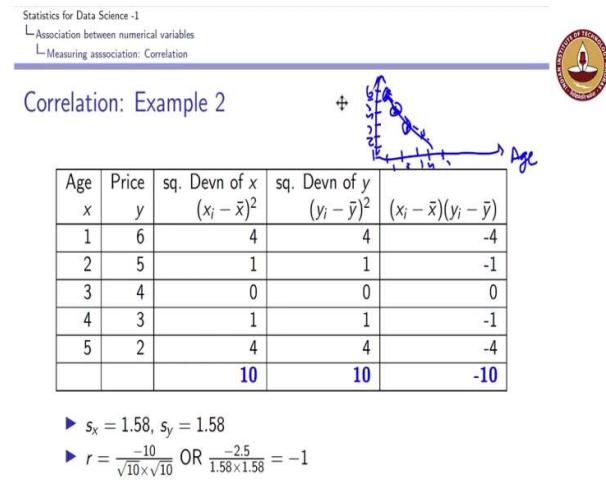
Now, let us go to the next example. In the next example, I looked at the age of a car and the price at which it is being sold. We know as cars get older, the price at which they are being sold comes down. So, again my price is recorded in lakhs of Indian rupees. So, as the age of the car, as my car gets older, my price comes down. Again, here I compute, so this was again a 3. So, you can go back this is what we have here, I had a 3, this the mean, so \bar{y} was 4, \bar{x} was 3, I compute the deviation, so I have a -2, -1, 0, 1, 2, and this is the square of the deviations which I have here.

Here again I have a 2, 1, 0, -1, -2, again the square is 4, 1, 0, -2^2 is 4 the product is -4, so this is a -4, -2×-2 , which is a 4, -1×1 , which is a -1. So, again, the numerator term, which is the sum of the product, cross products are equal to 10, then I have the square root of the sum of x deviations, which is equal to $\sqrt{10}$, the $\sqrt{(y_i - \bar{y})^2}$, this is again $\sqrt{10}$.

Hence, I have my first correlation metric which is $\frac{-1}{\sqrt{10}\sqrt{10}}$, which is -1, we already have computed the covariance as -2.5, which is $-10/4$ again recall, remember it is $-10/4$, when I am computing the sample covariance where, I am dividing it by $n - 1$ instead of the total number of observations. So, hence I am taking it to be a -2.5. The standard deviation is 1.58.

And I can see that $\frac{-2.5}{1.58 \times 1.58}$; I am going to get something which is close to -1.

(Refer Slide Time: 20:07)



So, we have already seen from the scatter plots of both of these, this has a negative correlation with age which is 1, 2, 3, 4, and 5; 3, 4, 5, 6. I have these as my points, 1, 2, 3, 4, 5, 6, 3 is a 4, 4 is a 3 and 5 is a 2. So, you can see that, as my car gets older, the price drops. And there is a perfect negative correlation between age and price, so I do not have to specify the units when I am talking about a correlation measure.

(Refer Slide Time: 20:46)

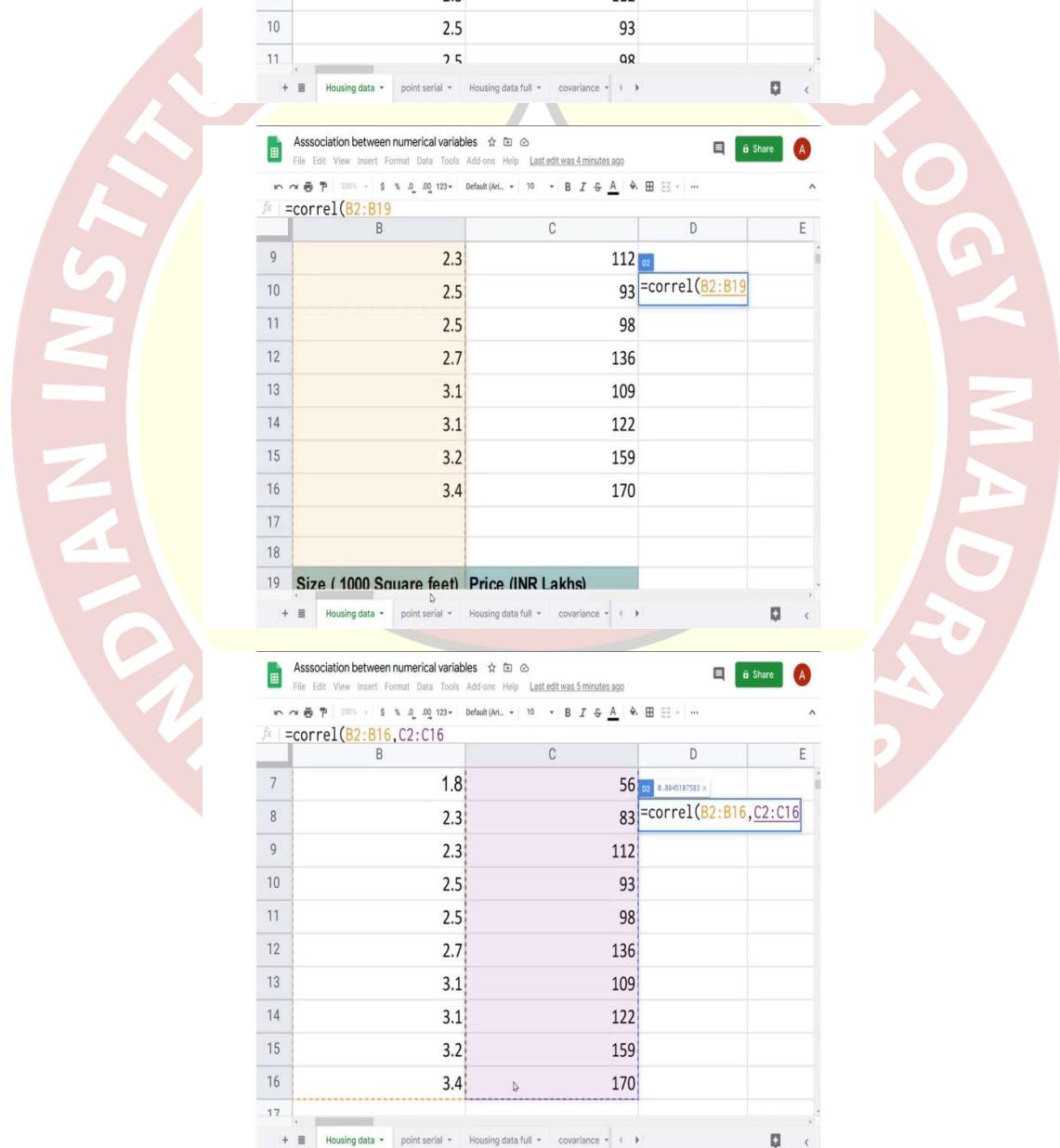


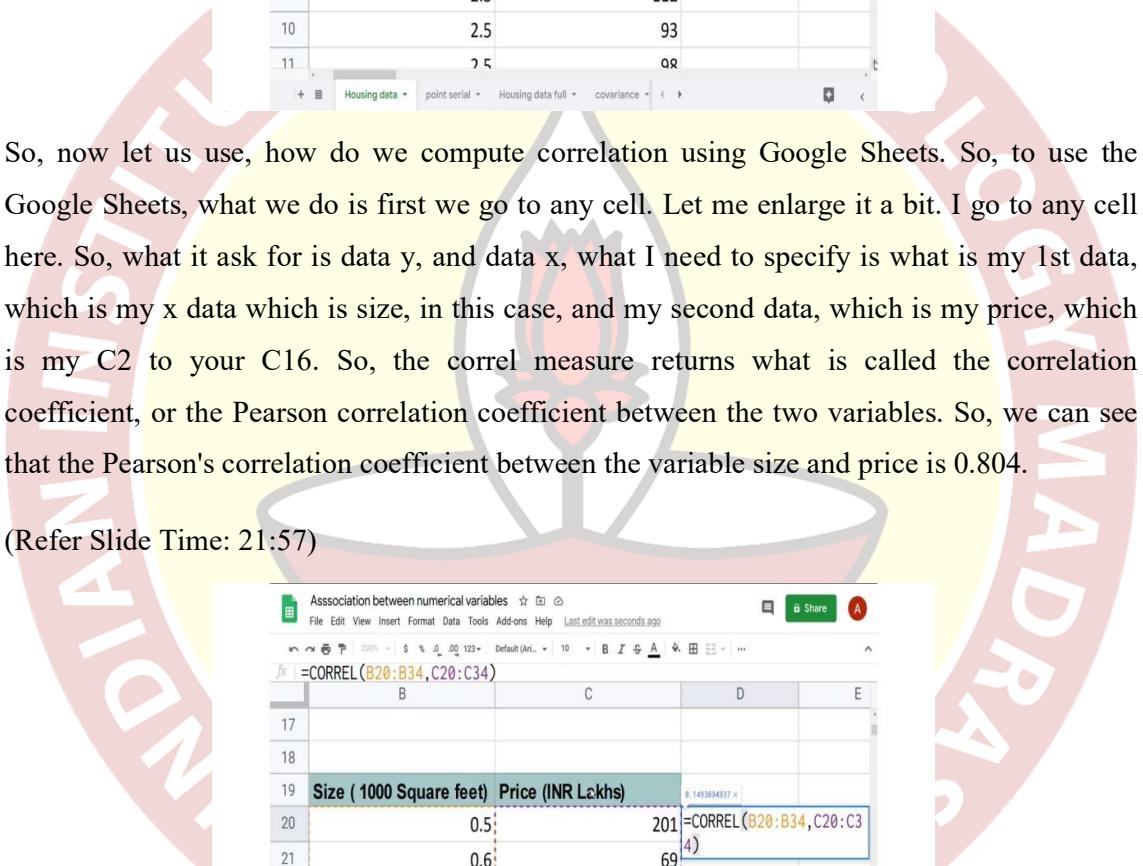
Correlation using google sheets

Step 1 The function CORREL(series1, series2) will return the value of correlation.

For example: If the data corresponding to x-variable (series1) is in cell A2:A6 and data corresponding to y-variable (series2) is in cells B2:B6; then CORREL(A2:A6,B2:B6) returns the value of the Pearson Correlation coefficient.





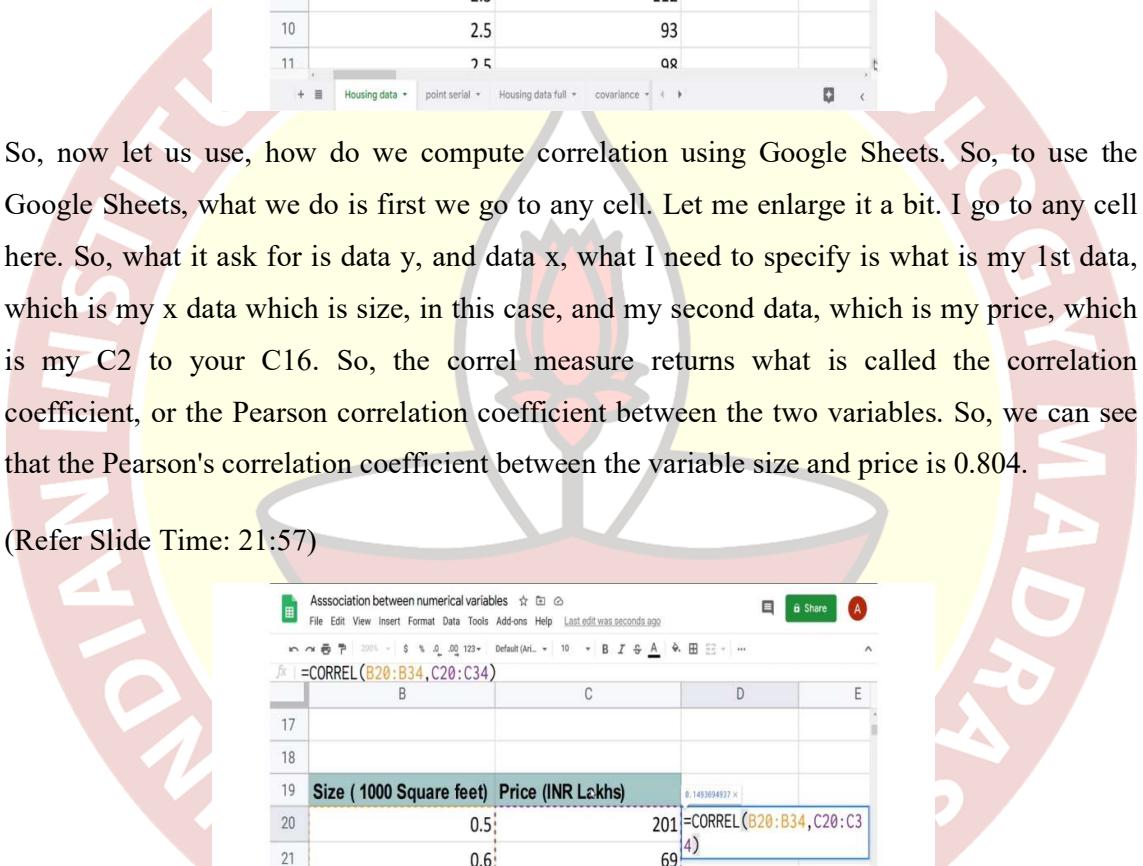


A screenshot of a Google Sheets document titled "Asssoication between numerical variables". The sheet contains a table with two columns: "Size (1000 Square feet)" and "Price (INR Lakhs)". The data starts from row 2 and ends at row 11. The formula bar shows the cell reference C2. The last cell in the table, C11, is selected.

	B	C	D	E
1	Size (1000 Square feet)	Price (INR Lakhs)		
2		0.8	68	0.8045107583
3		1	81	
4		1.1	72	
5		1.3	91	
6		1.6	87	
7		1.8	56	
8		2.3	83	
9		2.3	112	
10		2.5	93	
11		2.5	98	

So, now let us use, how do we compute correlation using Google Sheets. So, to use the Google Sheets, what we do is first we go to any cell. Let me enlarge it a bit. I go to any cell here. So, what it ask for is data y, and data x, what I need to specify is what is my 1st data, which is my x data which is size, in this case, and my second data, which is my price, which is my C2 to your C16. So, the correl measure returns what is called the correlation coefficient, or the Pearson correlation coefficient between the two variables. So, we can see that the Pearson's correlation coefficient between the variable size and price is 0.804.

(Refer Slide Time: 21:57)



A screenshot of a Google Sheets document titled "Asssoication between numerical variables". The formula bar shows the function =CORREL(B20:B34,C20:C34). The cell C20 is selected. The table below has two columns: "Size (1000 Square feet)" and "Price (INR Lakhs)". The data starts from row 20 and ends at row 27. The formula bar also shows the result 0.8045107583.

	B	C	D	E
17				
18				
19	Size (1000 Square feet)	Price (INR Lakhs)	0.8045107583	
20	0.5	201	=CORREL(B20:B34,C20:C34)	
21	0.6	69	0.8045107583	
22	0.9	122		
23	1.1	133		
24	1.3	207		
25	1.4	71		
26	1.5	149		
27	2	122		

Association between numerical variables

=CORREL(B20:B34,C20:C34)

	B	C	D	E
17				
18				
19	Size (1000 Square feet)	Price (INR Lakhs)		
20	0.5	201	=CORREL(B20:B34,C20:C34)	
21	0.6		0.1493694937	
22	0.9			
23	1.1			
24	1.3			
25	1.4			
26	1.5			
27	2			

Housing data point serial Housing data full covariance

Association between numerical variables

=CORREL(B20:B34,C20:C34)

	B	C	D	E
17				
18				
19	Size (1000 Square feet)	Price (INR Lakhs)		
20	0.5	201	0.1493694937	
21	0.6	69		
22	0.9	122		
23	1.1	133		
24	1.3	207		
25	1.4	71		
26	1.5	149		
27	2	122		

Housing data point serial Housing data full covariance

Similarly, let us look at the correlation coefficient between the next dataset. Again, what we had in that data set was a size and price for a different data set. Again, I find what is the correlation coefficient; my data here is B20 to B34, and C20 to C34, and I immediately see that the correlation coefficient is 0.149, which is very close to 0.

(Refer Slide Time: 22:29)

The image displays three separate screenshots of a Google Sheets document titled "Association between numerical variables".

Screenshot 1: A table with columns B, C, D, and E. Rows 37 through 47 show data points. Row 47 contains the formula `=corr1(B37:B50,C37:C50)`. The value in cell D47 is 1.330790126.

	B	C	D	E
37	3	8.8		
38	4	7.576		
39	5	6.49112		
40	5	5.5472744		
41	6	4.706128728		
42	6	3.974331993		
43	7	3.317668834		
44	7	2.746371886		
45	8	2.229343541		
46	9	1.75952888		
47	10	1.330790126		

Screenshot 2: A table with columns B, C, D, and E. Rows 30 through 36 show data points. Row 36 contains the header "Age of a car (years)" and "Price". Row 37 contains the formula `=CORREL(B37:B50,C37:C50)`. The value in cell D37 is -0.9271053621.

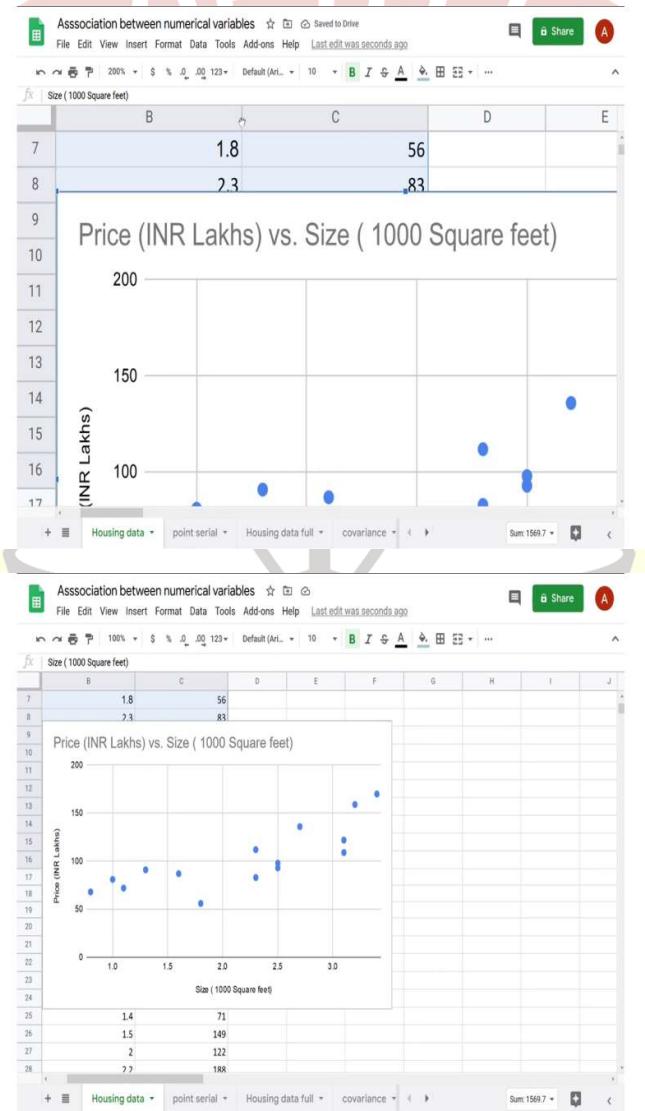
	B	C	D	E
30	2.7	88		
31	3	207		
32	3.1	133		
33	3.3	206		
34	3.4	90		
35				
36	Age of a car (years)	Price	-0.9271053621	
37	3	8.8	=CORREL(B37:B50,C37:C50)	
38	4	7.576		
39	5	6.49112		
40	5	5.5472744		

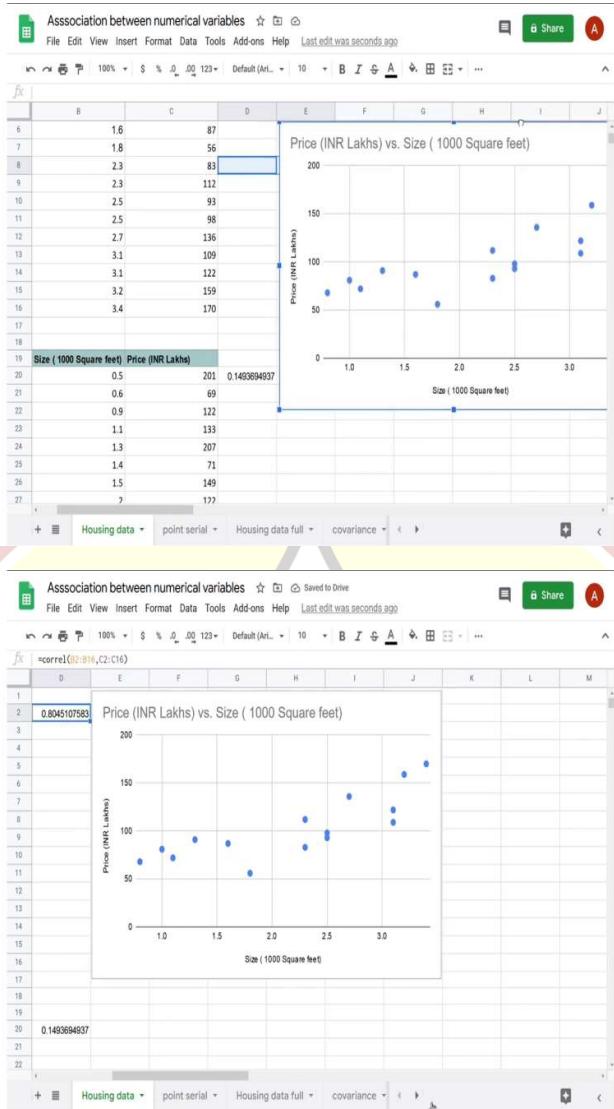
Screenshot 3: A table with columns B, C, D, and E. Rows 37 through 40 show data points. Row 37 contains the formula `=CORREL(B37:B50,C37:C50)`. The value in cell D37 is -0.9271053621.

	B	C	D	E
37	3	8.8	=CORREL(B37:B50,C37:C50)	
38	4	7.576		
39	5	6.49112		
40	5	5.5472744		

The third dataset was age of a car versus price of a car; I repeat the same here also. And I look at the correlation between the two variables here, which is going to be the age of a car and the price of a car. So this is B, so I just look at it, so let me look up to be B50 alone. And I notice that this correlation coefficient, the correl, the correl between these two is, what I am going to notice here, is given by minus. So, if you look at the data, you see it a - 0.927, which is also a large negative value close to - 1.

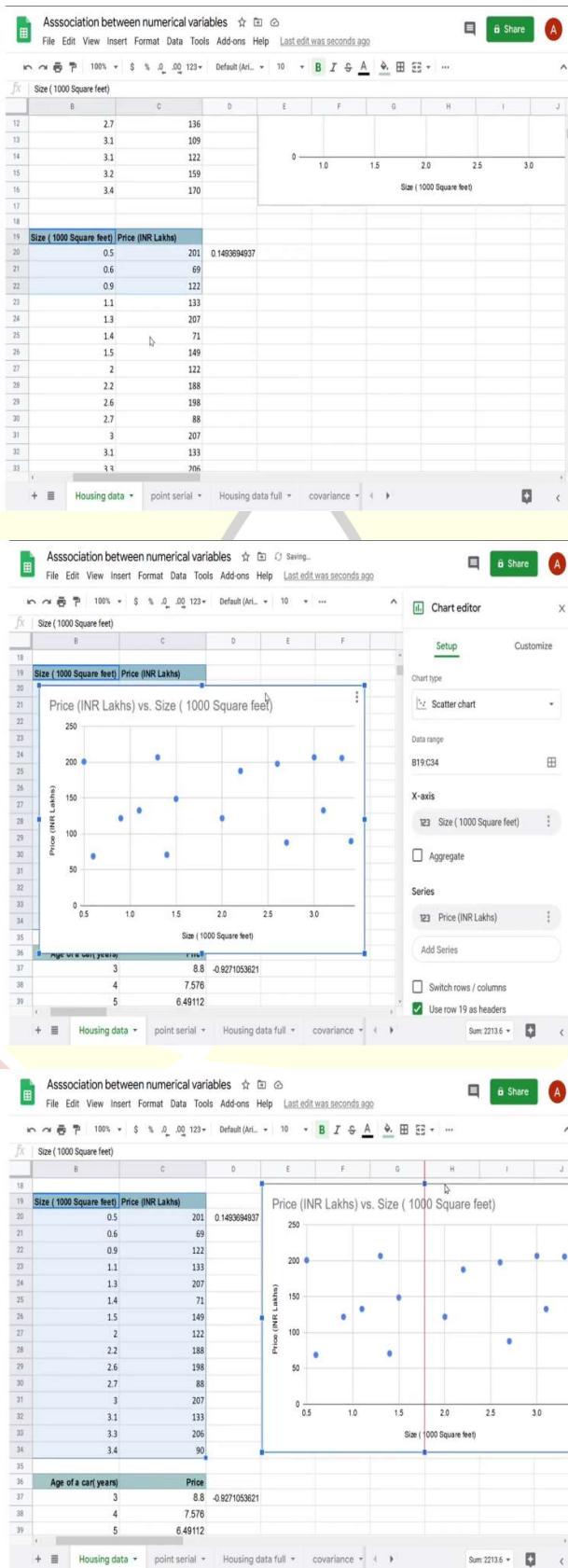
(Refer Slide Time: 23:27)

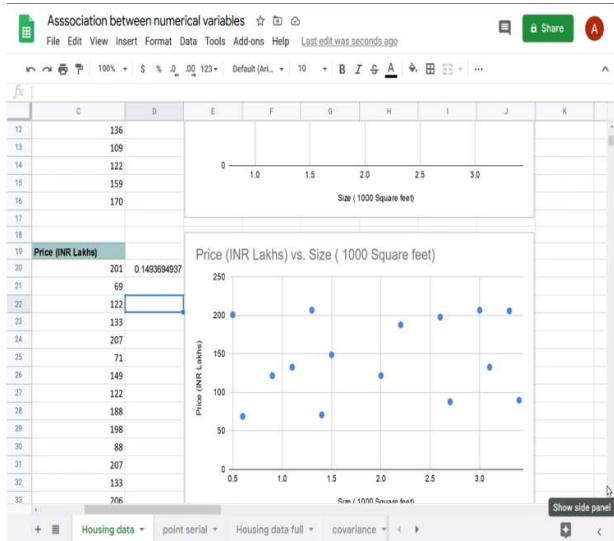




Now, one way to explain what is happening here is by looking at the scatter plots, so let me look at the scatter plot between my first dataset. So, again, I go, I plot a scatter plot between my first dataset. And, what I notice in my first dataset and the scatterplot is, there is a reasonable linear relationship between my x variable which was the size here, and my y variable, which was the price and this strength of this linear relationship is what is captured by my correlation, which is given by 0.804.

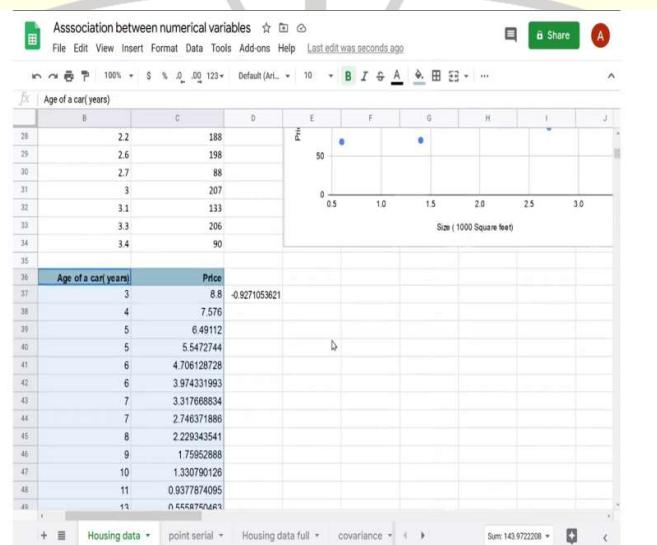
(Refer Slide Time: 24:11)

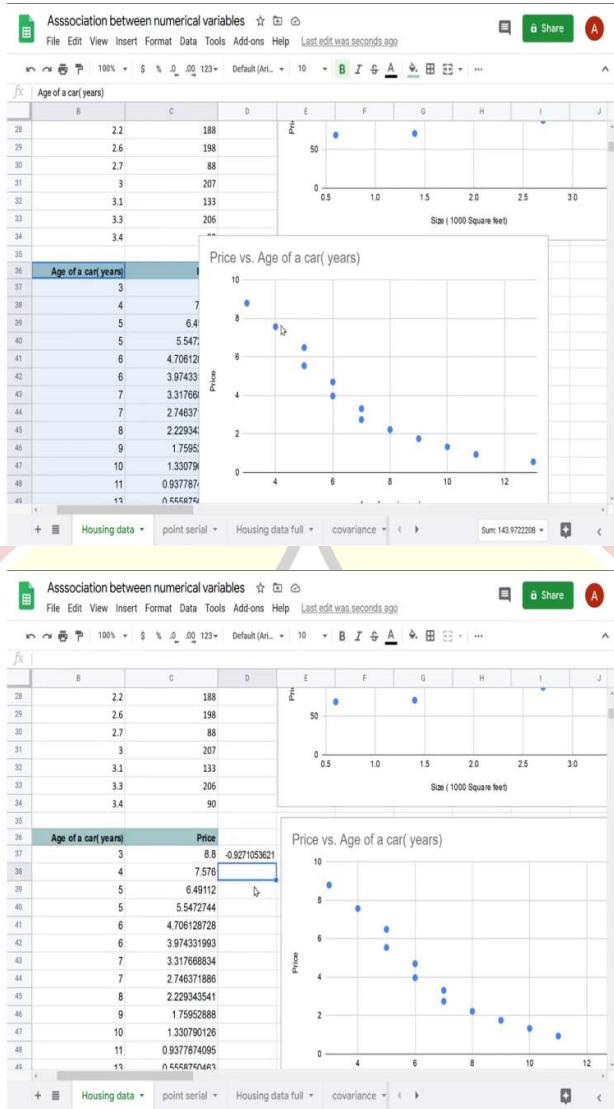




When I do the same scatterplot for the next dataset, I again go back and do the scatterplot for the next dataset. What I observe in this case is that even though the variables are the same, what we observe in this case is, the scatter which was, there was a pattern which was evident in the first dataset. I do not see that pattern here. And there is no pattern here, which is reflected in my correlation coefficient which is very low and close to 0, which is a 0.149.

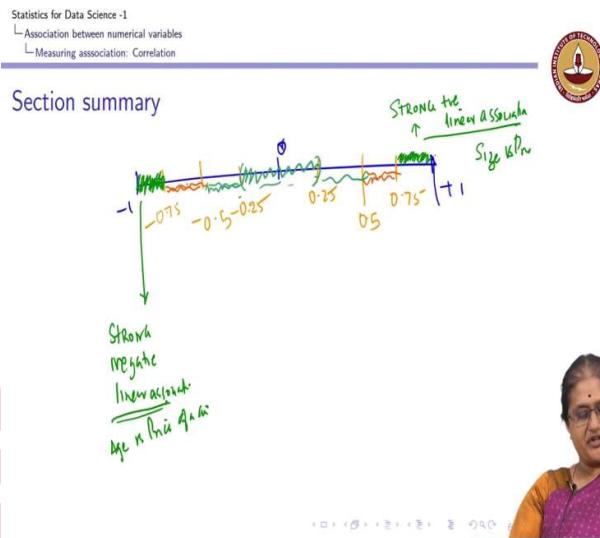
(Refer Slide Time: 24:53)





If I continue and do the same thing and plot a scatterplot in this scatterplot, I see a negative decrease in trend, and this quantifying or the strength of this negative relationship is quantified by this number -0.927. Hence, we can see that the correlation coefficient quantifies the strength of the linear relationship between my variables.

(Refer Slide Time: 25:27)



In other words, based on the correlation measure, I can tell the following that a the correlation measure lies between + 1 and - 1. So, I can have, so it can take any value. So, if it is this is a 0.5, this is a - 0.5, this could be a 0.75, - 0.75, this could be a - 0.25, this could be a plus 0.25, this could be a plus 0.75. So, I can start telling the following that if I have data, my correlation measure is between 0.75 and 1 then I could say that this indicates a strong positive linear association.

Similarly, if I have a correlation measure which is in this range, this could indicate a strong negative linear association. This association was my first example of size versus price of a home, this was the example of, this kind of relationship was age of a car versus price of a car, and depending, so this is a very strong, this portion could be just a portion where I have a reasonable positive and reasonable negative, this could have, this could indicate a weak, and between these two, this range could indicate no association. These are just indicator measures to say whether or to interpret; how strong is my linear relationship.

(Refer Slide Time: 27:42)

Statistics for Data Science -1
└ Association between numerical variables
 └ Measuring association: Correlation

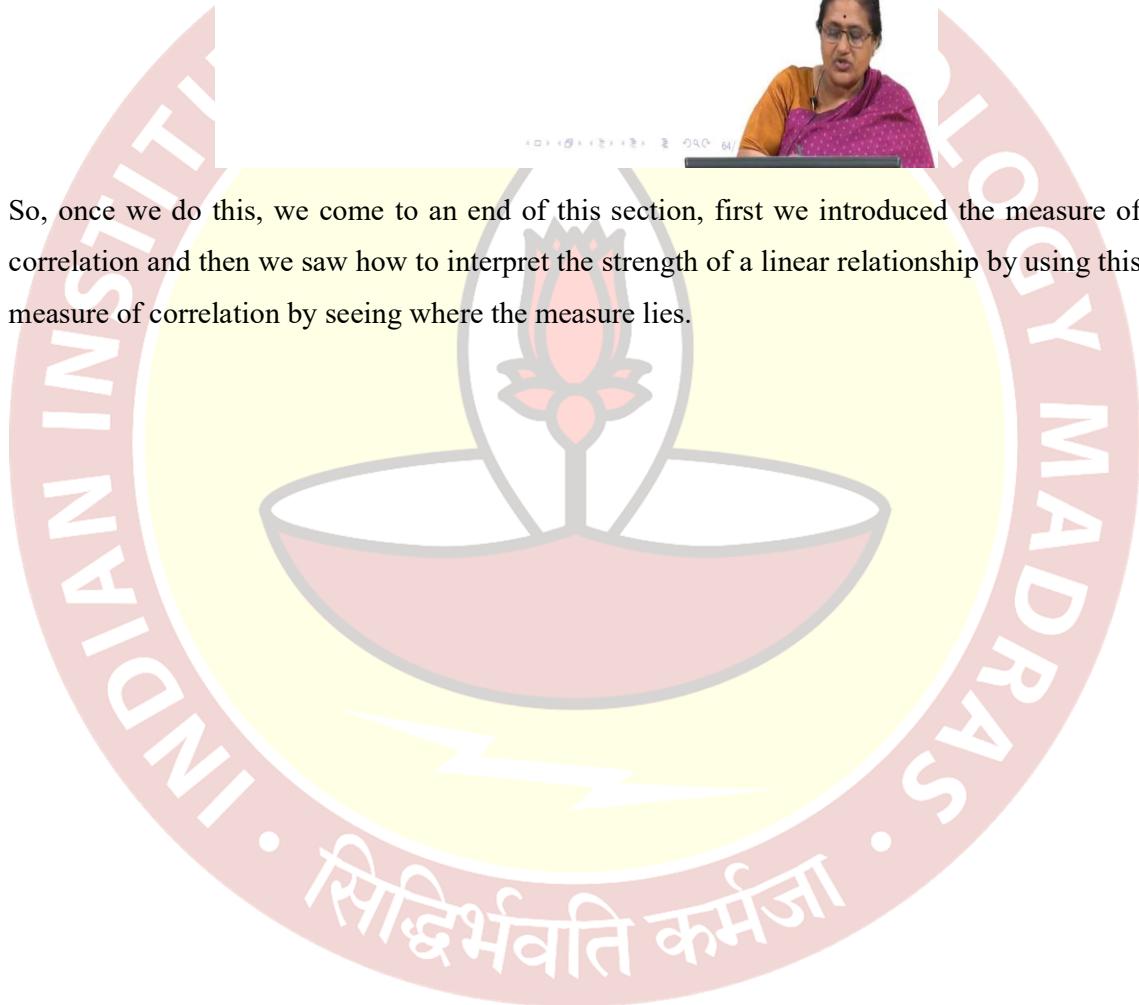


Section summary

1. Introduced measure of correlation.
2. Interpreting correlation between variables.



So, once we do this, we come to an end of this section, first we introduced the measure of correlation and then we saw how to interpret the strength of a linear relationship by using this measure of correlation by seeing where the measure lies.



Statistics for Data Science 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 4.8

Association between two numerical variables- Fitting a Line

(Refer Slide Time: 0:16)

Statistics for Data Science -1
└ Association between numerical variables
 └ Fitting a line



Learning objectives

1. Summarize the linear association between two variables using the equation of a line.
2. Understand the significance of R^2



Navigation icons: back, forward, search, etc.

So, we have seen that this correlation and covariance are measures of linear relationship or linear association between two numerical variables; it measures a strength of a association.

(Refer Slide Time: 0:34)

Statistics for Data Science -1
└ Association between numerical variables
 └ Fitting a line

Learning objectives



1. Summarize the linear association between two variables using the equation of a line.
2. Understand the significance of R^2

Navigation icons: back, forward, search, etc.

So, when I say that the association is a linear association. The next natural question to ask is, can I summarize this linear association through a mathematical equation in particular? The question we are asking as can this linear relationship be summarized through a equation of a line.

(Refer Slide Time: 1:03)

Statistics for Data Science -I

↳ Association between numerical variables

↳ Fitting a line

Summarizing the association with a line

(x, y)

- ▶ The strength of linear association between the variables was measured using the measures of Covariance and Correlation.

So, the question we are asking here is, when I am saying that x and y are numerical variables and they have association which I expect to be linear.

(Refer Slide Time: 1:15)

(Refer Slide Time: 1:15)

INDIA

STATISTICS FOR DATA SCIENCE - I

Association between numerical variables

Fitting a line

Summarizing the association with a line

STATISTICAL INSTITUTE OF TECHNOLOGY
MADRAS

- The strength of linear association between the variables was measured using the measures of Covariance and Correlation.
- The linear association can be described using the equation of a line. $y = mx + c$

best line of fit

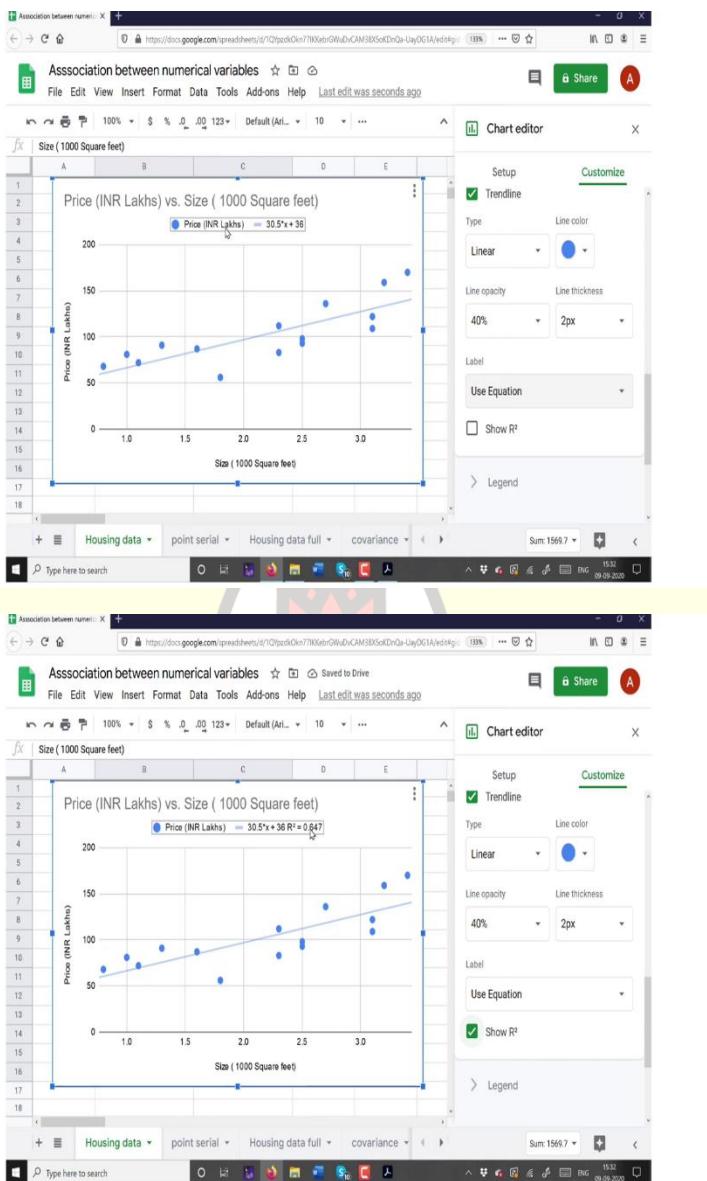
Then the next question we are asking is can I described that association using the equation of a line? If yes, then how do I compute this equation of a line? The mathematics behind coming up with an equation of a line is beyond the scope of this particular course. You will be learning about how to come out with what we call the best line of fit in your future courses. But, what I want you to see at this point of time is, yes; the relationship can be summarized. How do I summarize this?

(Refer Slide Time: 2:06)

	A	B	C	D	E	F	G	H	I
1		Size (1000 Square feet)	Price (INR Lakhs)						
2	1	0.8	68	0.8045107583					
3	2	1	81						
4	3	1.1	72						
5	4	1.3	91						
6	5	1.6	87						
7		1.8	56						
8	7	2.3	83						
9	8	2.3	112						
10	9	2.5	93						
11	10	2.5	98						
12	11	2.7	138						
13	12	3.1	109						
14	13	3.1	122						
15	14	3.2	159						
16	15	3.4	170						
17									
18									

So, let us go back to our Goggle sheets; so you can see that I start with my first data here, which is my size wise is the price. I go to this data set. So, how do I find the equation of a line using a Google sheet?

(Refer Slide Time: 2:28)



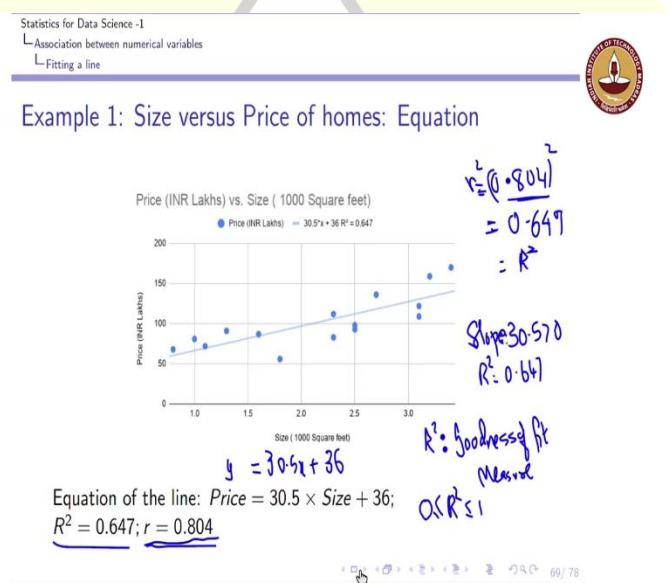
So, the first step is you open the scatter plot. So, I first go and I insert a scatter plot and this scatter plot is what I have here; so first I go and I insert this scatter plot. So, the next step is under the Customize tab, so I have a scatter plot here. This is scatter plot of size versus price; size is my explanatory variable, price is my response variable. Under the Customize tab click on Series, so under the Customize tab I click on series here; now within this click on trendline.

I clicked on series, I clicked on trendline; so, you can see a blue line which passes through my points. Now, the next thing is this trendline, so the question we ask is can I capture this linear relationship using the equation of a line? I can see that I can fit a line through the points passing

through the points. Now, further if I want to know what is the equation under the Label tab; so within this I have a Label tab. Under this Label tab click on this Use Equation; so it gives me the equation. You can see here the equation states price to equal $30.5 \times x$; x here is my size +36.

This is of the type $y = mx + c$, which all of us know is the equation of a straight line. Now, further I can also ask it to report this R^2 ; so it gives an R^2 of 0.647. What is this R^2 capture? This R^2 actually captures the proportion of variance in my data set; that is captured by this line. Again to go into the mathematics of this R^2 and how to derive it is beyond the scope of this course; but, R^2 basically is also a measure of how good a fit is this line.

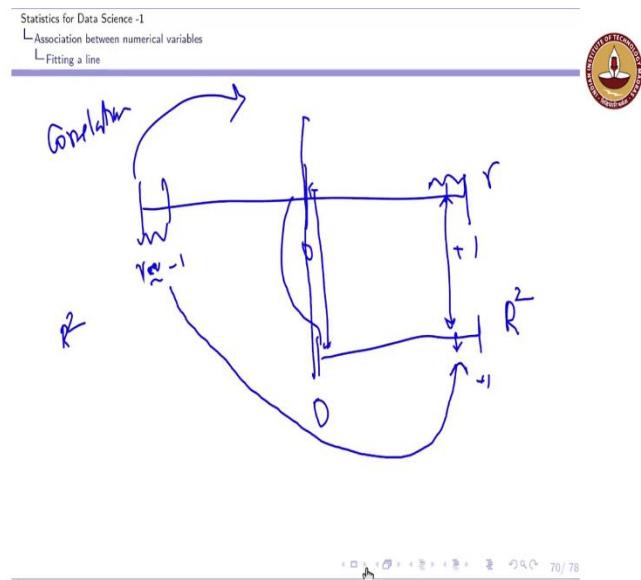
(Refer Slide Time: 5:01)



This R^2 takes the value, so this R^2 is also referred to a goodness of fit measure. It is an R^2 , so this takes values between 0 and 1. The closer it is equal to 1, says that my fit is a good fit; the closer it is to 0, tells my fit is not a good fit for my data. So, now let us look at what is this R^2 for different examples which we have looked. In the first example this is what I have demonstrated just now, I have price which is my y ; the response variable is $30.5 \times x + 36$. So, the slope of this line is 30.5, you can see that the slope is positive; and your R^2 is 0.647.

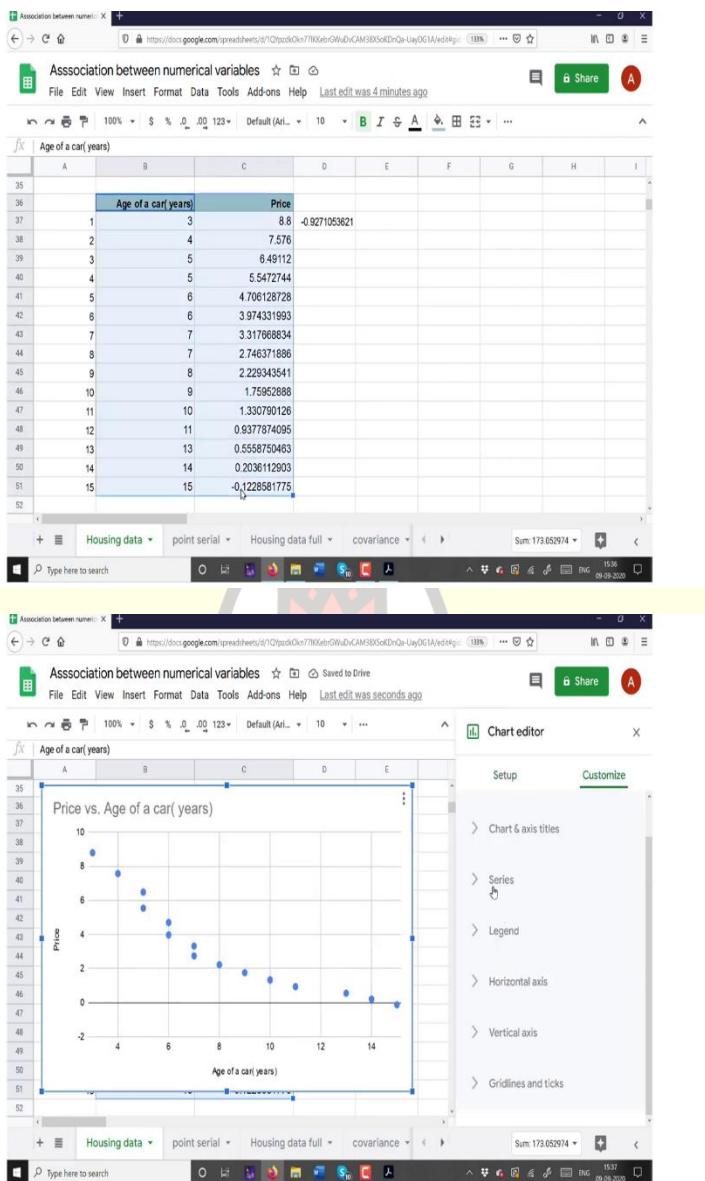
Recall my correlation coefficient was 0.804. An interesting observation is if I square this correlation coefficient; so you can see that I computed the correlation coefficient of my first data set to be 0.804. If I just square this term, I get 0.647 which is precisely this value of my R^2 .

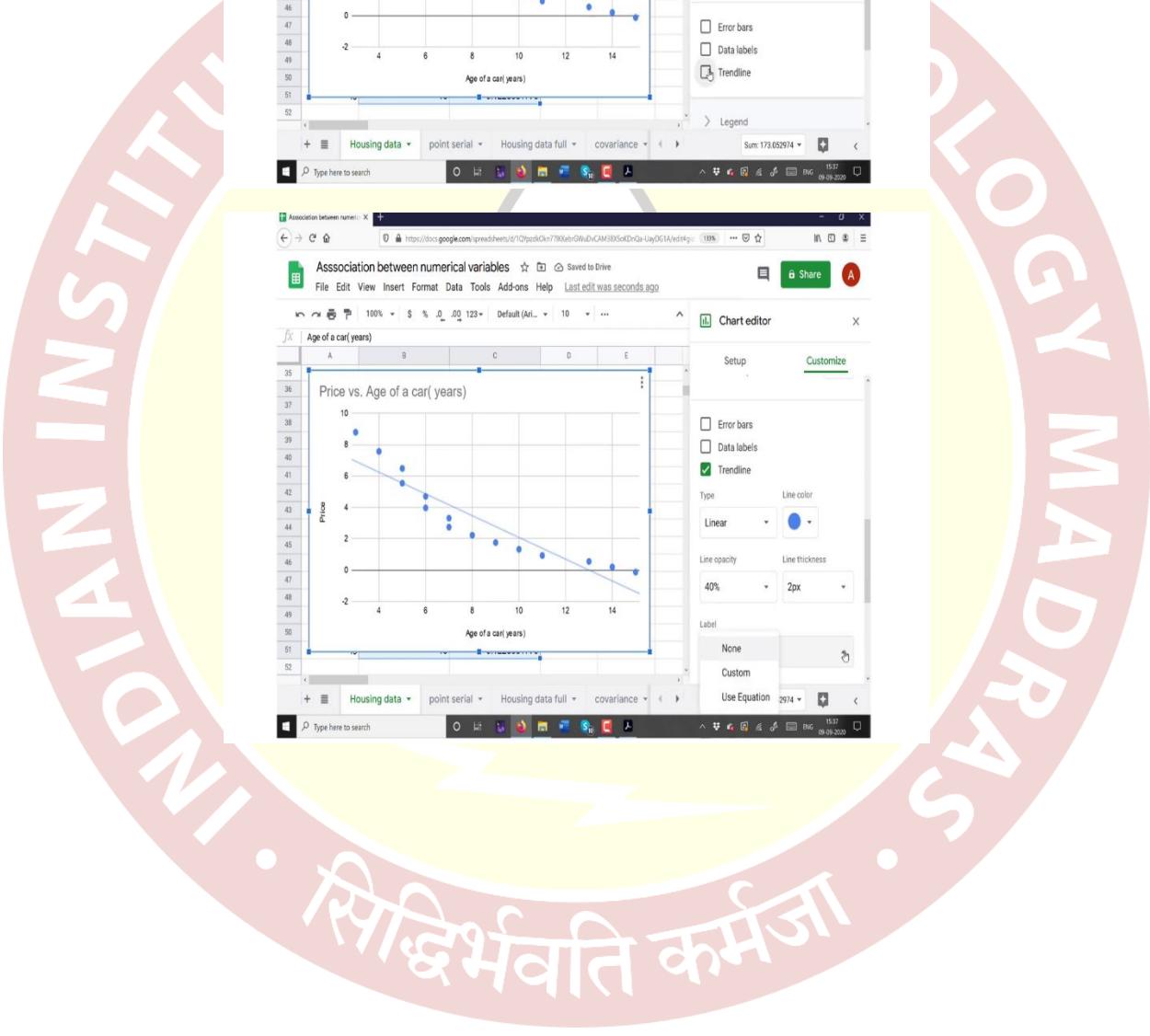
(Refer Slide Time: 6:45)

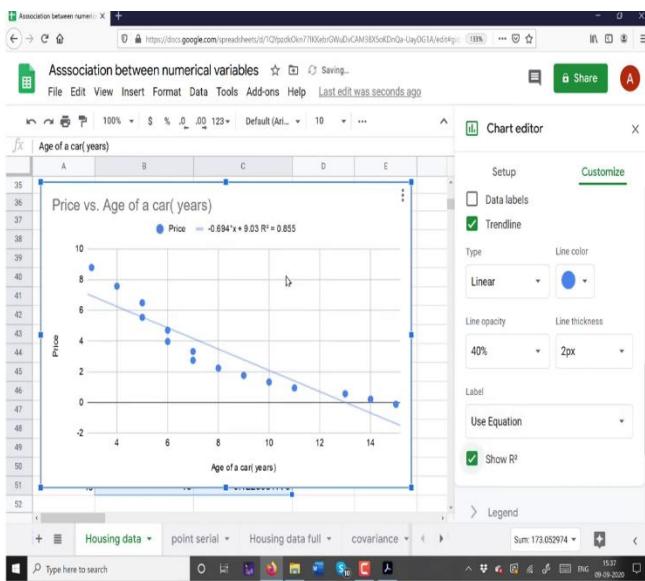


So, let us go back to our data here, so you can see that when I have a correlation which is closer to 1; my R^2 which is the square of this would also be closer to 1. This is about r , the first thing is my R^2 will only lie between 0 and 1; so, in a sense you can imagine to flip this along this 0. So, if I have a value which is here r value here; my R^2 is going to be closer to 1. So, as my r goes to 0, my R^2 will also tend to 0.

(Refer Slide Time: 7:35)

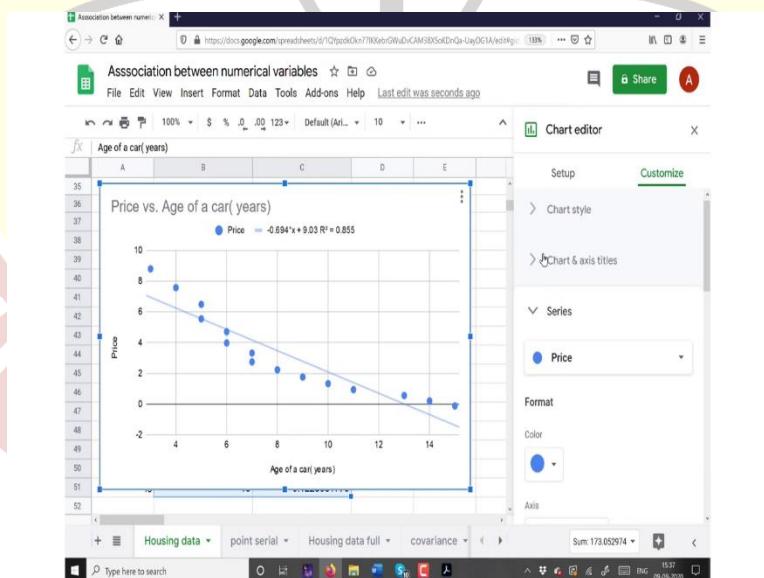






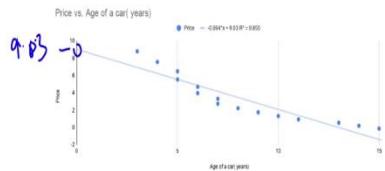
Let us look at an example to illustrate what we have just discussed. So, if I go and look at the same thing for this data set, we know about the age of a car and the price of a car. This is the next data set. Again I go and I click on a scatter plot, again I go to my Customize option; under series I plotted a trendline. And again I ask for a equation with a R^2 .

(Refer Slide Time: 8:14)



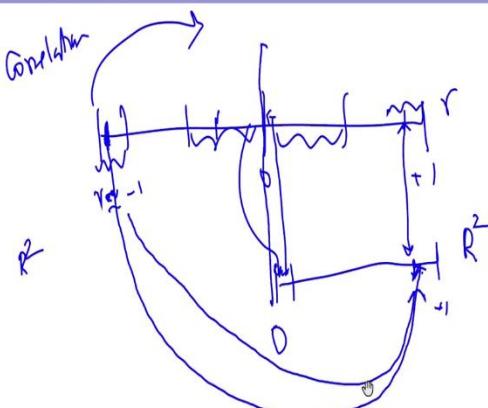


Example 2: Age versus Price of cars: Equation



Equation of the line: $\text{Price} = -0.694 \times \text{Age} + 9.03$;
 $R^2 = 0.855$; $r = -0.9247$

71/78



70/78

So, here what you noticed? You notice that the price is $-0.694x + 3$. With again an R^2 of 0.855; which is price equal to -0.694 . So, my slope of the line is negative, which is 0.694; the intercept is 9.03. And then we also see that the R^2 is 0.855; even though your r was -0.92 . So, my r in this case or my correlation in this case was actually here; the correlation was closer to -1 . It was very strong negative linear relationship; my R^2 is closer 1.

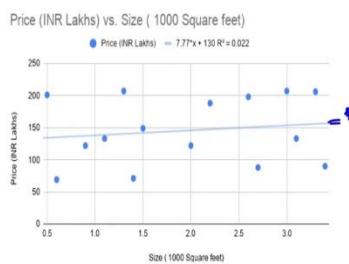
So, you can see that this goodness of fit measure takes a values between 0; and 1 but it does not tell me the direction of the relationship. For the direction, whether it is a positive relationship or negative relationship. I either look at the correlation coefficient or the sine of the slope; in this case the slope is a negative number.

(Refer Slide Time: 9:40)

Statistics for Data Science -I
└ Association between numerical variables
 └ Fitting a line



Example 3: Size versus Price of homes: Equation



Equation of the line: $Price = 7.77 \times Size + 130$;
 $R^2 = 0.022$; $r = 0.149$.

Now, let us look at the third example. The third example my r was very negligible; so it says that almost no relation. It is giving me the equation of a line; but what you have noticed about this line, it is almost parallel to my x axis. But, and your R^2 is closer to 0. So, when your R^2 is closer to 0; the goodness of the fit measure is also very low. So, when your R^2 is close to 0; it quantifies the strength of the relationship. And you can say that my line is not actually describing this association very well.

(Refer Slide Time: 10:26)

Statistics for Data Science -1
└ Association between numerical variables
 └ Fitting a line

Section summary

$y = mx + c$

Response = $\underset{m > 0}{\underset{m < 0}{\text{Explanatory}}} + c$

1. Equation of a line describing linear relationship between two variables.
2. Interpreting slope, R^2 of the line.

Association between Categorical variables

- 1. Contingency table
- 2. Relative frequency

Association between Numerical variables

- 1. Scatter plot - visual direction
- 2. Covariance Correlation
- 3. Equation of line, R^2 .



So, in summary what we have looked that is how do we obtain the equation of a line; we are not again gone into the mathematics of trying to find, what is a line of best fit. However, we got this equation of a line through our Google sheets. And when we looked at the equation of a line; so I get it as $y = mx + c$, y is my response variable, x is my explanatory variable.

And m is the slope of the line, c is the intercept. The sign of the slope, so if $m > 0$; then this says that I have a positive relationship. $M < 0$, it says that I have a negative relationship; and we also define what is R^2 , which is a goodness of fit measure. R^2 lies between 0 and 1 and it is closer to 1; we are getting equations of the line. It says that the fit, I am talking about when it comes to a fit; I mean that the line is not capturing the variability in the data as much as in the other case. So, the proportion of variability in my data set that is captured by this line is very low, if R^2 is closer to 0. And it is pretty high, if the $R^2 = 1$.

So, with this we actually have seen the following. In this section we started by looking at the association between two numerical variables; earlier we looked at association between categorical variables. Now, in this case the key thing is, we first learned about how we set up what we called is contingency table. Here we started with a scatter plot to look at the association, and then we introduced the notion of relative frequency here.

We looked at row relative frequency and column related frequency. If they are the same for all rows and columns, then after we set. When we looked at association between numerical variables, we started by looking at a scatter plot. From here we wanted to just have a visual inspection, within the visual inspection we identified what was the direction.

Whether it is a positive trend or a negative trend, whether it is a curve or a line, whether there they are tight, whether they clustered, or the presence of outliers; these are the four things which we looked at. We finally then we said that okay I am focus on a linear association between variables.

Now, if I want to know the strength of this association; I introduced two main numerical measures, which are covariance and correlation measures. And finally we also looked at how to summarize or describe this linear relation through a equation of a line. We introduced the concept of R^2 which is nothing but the goodness of the fit measure.

(Refer Slide Time: 14:16)

Statistics for Data Science -1
└ Association between numerical variables
 └ Fitting a line

Section summary

Variable

```

graph TD
    Variable --> Categorical
    Variable --> Numerical
  
```

Categorical

Numerical

<u>Association between categorical</u> <ul style="list-style-type: none"> 1. Contingency table 2. Relative frequencies 	<u>Association between Numerical</u> <ul style="list-style-type: none"> 1. Scatter plot - visual direction, Curve-linear, Tight, Outliers 2. Covariance Correlation 3. Equation of line, R^2
--	--

So, now if you look at variables, again you go back to your where we started from. And we saw that when we look at variables; I can broadly classify my variables or my data as categorical data and I can numerical data. So, this portion looked at association when both my variables or pair of variables are categorical. This looked at what happens when pair of variables are numerical in nature.

Statistics for Data Science - 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture 4.9
Association between categorical and numerical variables

(Refer Slide Time: 0:17)

Statistics for Data Science - 1
↳ Association between categorical and numerical variable

Introduction X, Y

▶ Understand the association between a categorical variable and numerical variable.

▶ Assume the categorical variable has two categories (dichotomous)

gender $\begin{cases} \text{MALE} \\ \text{FEMALE} \end{cases}$

income $\begin{cases} \text{HIGH} \\ \text{LOW} \end{cases}$

So, in this portion we are going to understand how to capture the association between a numerical and a categorical variable. Here I am assuming my categorical variable has only two categories, in other words if I am looking at gender for example, it has two categories male and female. I could have another variable which is say income, I could just look at two categories which is just high category and low category.

So, I am coming my categorical variable has two (variables) two categories and this is referred to as a dichotomous variable. So, now we are going to see how we summarize the association or understand the association between a categorical variable and a numerical variable. Let us look at the following example.

(Refer Slide Time: 1:17)

Statistics for Data Science -1
↳ Association between categorical and numerical variable



Example 1: Gender versus marks

A teacher was interested in knowing if female students performed better than male students in her class. She collected data from twenty students and the marks they obtained on 100 in the subject.



Statistics for Data Science -1
↳ Association between categorical and numerical variable



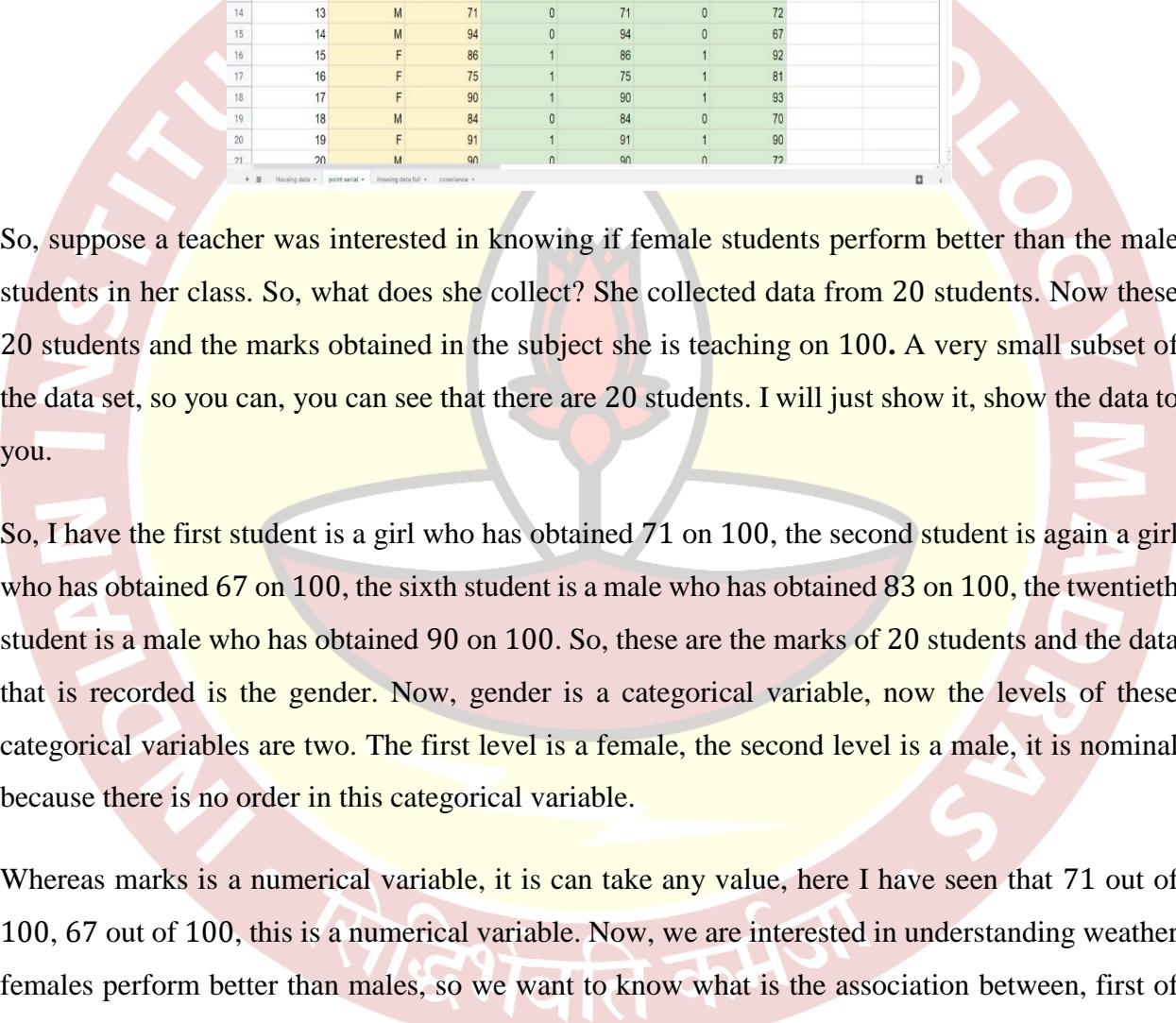
Example 1: Gender versus marks-Data

	Gender	Marks
1	F	71
2	F	67
3	F	65
4	M	69
5	M	75
6	M	83
7	F	91
8	F	85
9	F	69
10	F	75
11	M	92
12	F	79
13	M	71
14	M	94
15	F	86
16	F	75
17	F	90
18	M	84
19	F	91
20	M	90



Statistics for Data Science -1
↳ Association between categorical and numerical variable





A screenshot of a Microsoft Excel spreadsheet titled "Association between numerical variables". The data is organized into two main sections: "Gender-coded" and "Marks".

	B	C	D	E	F	G	H	I
1	Gender	Marks	Gender-coded	Marks	Gender-coded	Marks		
2	1	F	71	1	71	1	86	#DIV/0!
3	2	F	67	1	67	1	92	
4	3	F	65	1	65	1	92	
5	4	M	69	0	69	0	68	
6	5	M	75	0	75	0	70	
7	6	M	83	0	83	0	66	
8	7	F	91	1	91	1	91	
9	8	F	85	1	85	1	90	
10	9	F	69	1	69	1	90	
11	10	F	75	1	75	1	89	
12	11	M	92	0	92	0	68	
13	12	F	79	1	79	1	92	
14	13	M	71	0	71	0	72	
15	14	M	94	0	94	0	67	
16	15	F	86	1	86	1	92	
17	16	F	75	1	75	1	81	
18	17	F	90	1	90	1	93	
19	18	M	84	0	84	0	70	
20	19	F	91	1	91	1	90	
21	20	M	90	0	90	0	72	

So, suppose a teacher was interested in knowing if female students perform better than the male students in her class. So, what does she collect? She collected data from 20 students. Now these 20 students and the marks obtained in the subject she is teaching on 100. A very small subset of the data set, so you can, you can see that there are 20 students. I will just show it, show the data to you.

So, I have the first student is a girl who has obtained 71 on 100, the second student is again a girl who has obtained 67 on 100, the sixth student is a male who has obtained 83 on 100, the twentieth student is a male who has obtained 90 on 100. So, these are the marks of 20 students and the data that is recorded is the gender. Now, gender is a categorical variable, now the levels of these categorical variables are two. The first level is a female, the second level is a male, it is nominal because there is no order in this categorical variable.

Whereas marks is a numerical variable, it is can take any value, here I have seen that 71 out of 100, 67 out of 100, this is a numerical variable. Now, we are interested in understanding whether females perform better than males, so we want to know what is the association between, first of all we are asking the question, is there an association between gender and the marks, or are they not associated with each other.

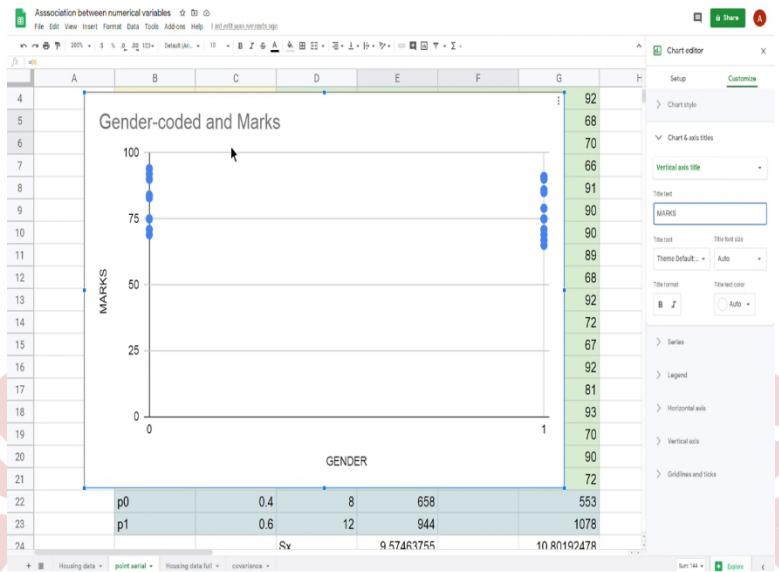
We introduced a concept which is referred to as a point bi-serial correlation measure. So, let us start by looking at a scatter of this data. To look at a scatter of this data what I first do is the

following, I code my categorical variable. What do I mean by coding my categorical variable? I have gender female and male here so I am just coding it, because now I am looking at whether I can quantify the correlation between these two variables. So if I just write for example, if I just try and see I will put a correlation measure and I, you can see that it returns an error because it says that correlation the (valid), it has no valid input data.

It has no valid input data because this is categorical and it is text. So, let me code this variable. So, this coding I am just arbitrarily choosing female as 1 and male as 0, so you can see that this data is coded in the following form. So, here I have two data sets, this could be the data obtained in one test and this could be the data or this could be the marks obtained in of the same 20 people in one test and this could be the marks obtained of the same 20 people in another test.

(Refer Slide Time: 4:52)

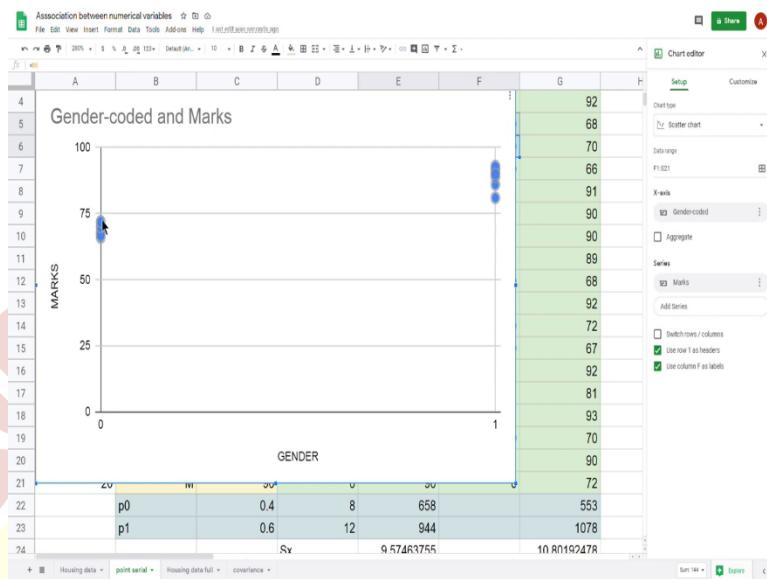
	A	B	C	D	E	F	G	H	I
1		Gender	Marks	Gender-coded	Marks	Gender-coded	Marks		
2	1	F	71	1	71	1	86		
3	2	F	67	1	67	1	92		
4	3	F	65	1	65	1	92		
5	4	M	69	0	69	0	68		
6	5	M	75	0	75	0	70		
7	6	M	83	0	83	0	66		
8	7	F	91	1	91	1	91		
9	8	F	85	1	85	1	90		
10	9	F	69	1	69	1	90		
11	10	F	75	1	75	1	89		
12	11	M	92	0	92	0	68		
13	12	F	79	1	79	1	92		
14	13	M	71	0	71	0	72		
15	14	M	94	0	94	0	67		
16	15	F	86	1	86	1	92		
17	16	F	75	1	75	1	81		
18	17	F	90	1	90	1	93		
19	18	M	84	0	84	0	70		
20	19	F	91	1	91	1	90		
21	20	M	90	0	90	0	72		



So, let us first start with a scatter plot of this data. So, again I go and I plot a scatter plot. I do not want a line chart, I want a scatter plot. And within the scatter plot I am using my column D as labels because I, column D is taking only two values, I am using column D as labels. And within this column D again what I do is, I go and I customize it further by looking at the major spacing time which is step and I am putting a step value of 1. So, you can see that my, I have constructed my scatter plot which, in which I have my X , I can go back to my titles, the title is gender coded versus marks.

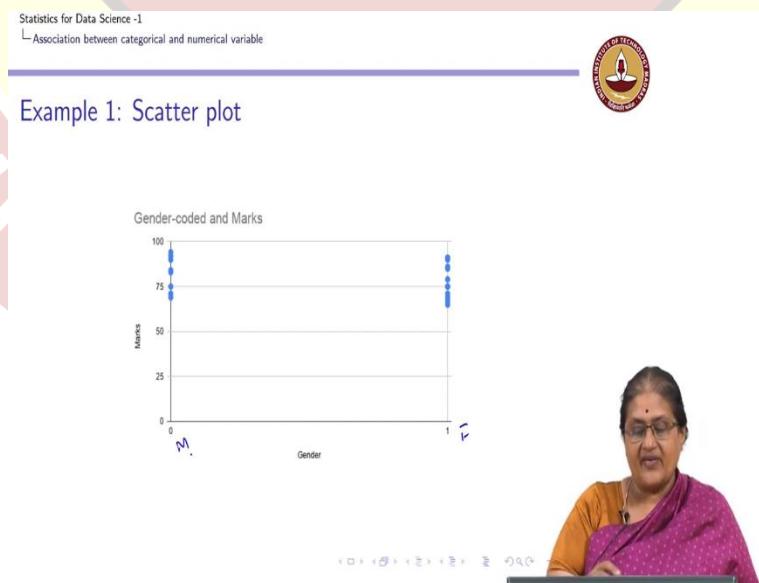
My horizontal axis I have gender and on my vertical axis I have marks out of 100. And the way I can interpret this scatter plot is, you can see that this 0, 0 represents a male here and 1 represents a female here. So you can see that whether it is male or female all of them have obtained marks in the same range. So, the scatter plot tells us an important story here and the story which the scatter plot, this scatter plot tells us is irrespective of gender the distribution of marks seems to be the same, because both female and male seem to have performed equally well in a particular range. But now let us look at the second dataset.

(Refer Slide Time: 7:00)



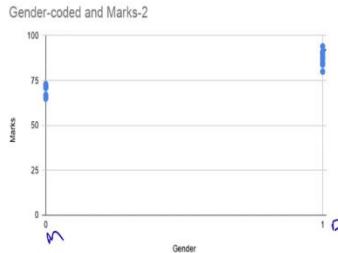
So, if I go back and change this data range, let me look at the data range taking so, it is sorry F21 to F1 to G21 that is a second data range of my data. So, for this data range F1 to G21. Now, you see that there is some difference. What is difference you notice here? You can see that 0 or the males are clustered in this region and females are clustered in this region to indicate females have performed better, they have obtained higher marks than the males in general.

(Refer Slide Time: 7:47) 20:03, 22:11





Example 1: Scatter plot



So, you see that in the earlier case, so you can see that in the earlier case that is in this case, the first case I see that 1 is again female. 0 is a male. And you can see that the marks obtained are the same. Whereas in the second case again I can see that female seem to have obtained higher marks than the males in general. This is a distribution of marks for males and the females. So, the question is, again here I do not have a line, I have just these points. Fitting a line for this kind of data is of no use to me, so how do I summarize the strength of this association.

(Refer Slide Time: 8:35)

X: Marks Female

Point Bi-serial Correlation Coefficient

- Let X be a numerical variable and Y be a categorical variable with two categories (a dichotomous variable).
 - The following steps are used for calculating the Point Bi-serial correlation between these two variables:

Step 1 Group the data into two sets based on the value of the dichotomous variable Y . That is, assume that the value of Y is either 0 or 1.

Step 2 Calculate the mean values of two groups: Let \bar{Y}_0 and \bar{Y}_1 be the mean values of groups with $Y = 0$, and $Y = 1$, respectively.

Step 3 Let p_0 and p_1 be the proportion of observations in a group with $Y = 0$ and $Y = 1$, respectively, and s_X be the standard deviation of the random variable X .

The correlation coefficient

$$r_{pb} = \left(\frac{\bar{Y}_0 - \bar{Y}_1}{S_x} \right) \sqrt{p_0 p_1}$$

X: Marks
Y: Gender

$$Y = \begin{cases} 0 & \text{Male} \\ 1 & \text{Female} \end{cases}$$



Point Bi-serial Correlation Coefficient

► Let X be a numerical variable and Y be a categorical variable with two categories (a dichotomous variable).

► The following steps are used for calculating the Point Bi-serial correlation between these two variables:

Step 1 Group the data into two sets based on the value of the dichotomous variable Y. That is, assume that the value of Y is either 0 or 1.

Step 2 Calculate the mean values of two groups: Let \bar{Y}_0 and \bar{Y}_1 be the mean values of groups with $Y = 0$, and $Y = 1$, respectively.

Step 3 Let p_0 and p_1 be the proportion of observations in a group with $Y = 0$ and $Y = 1$, respectively, and s_x be the standard deviation of the random variable X.

The correlation coefficient

$$r_{pb} = \left(\frac{\bar{Y}_0 - \bar{Y}_1}{s_x} \right) \sqrt{p_0 p_1}$$

n : Total no. of obs.
 n_0 : No. of obs. "0" am
 n_1 : No. of obs. "1"

$$\sqrt{\frac{n_0 \cdot n_1}{n-1}}$$



Association between numerical variable

	A	B	C	D	E	F	G	H	I
1	Gender	Marks	Gender-coded	Marks	Gender-coded	Marks			
2	1	F	71	1	71	1	86	86	
3	2	F	67	1	67	1	92	=71+67+65+91+85+69+75+79+86+75+90+91	
4	3	F	65	1	65	1	92	=71+67+65+91+85+69+75+79+86+75+90+91	
5	4	M	69	0	69	0	68		
6	5	M	75	0	75	0	70		
7	6	M	83	0	83	0	66		
8	7	F	91	1	91	1	91		
9	8	F	85	1	85	1	90		
10	9	F	69	1	69	1	90		
11	10	F	75	1	75	1	89		
12	11	M	92	0	92	0	68		
13	12	F	79	1	79	1	92		
14	13	M	71	0	71	0	72		
15	14	M	94	0	94	0	67		
16	15	F	86	1	86	1	92		
17	16	F	75	1	75	1	81		
18	17	F	90	1	90	1	93		
19	18	M	84	0	84	0	70		
20	19	F	91	1	91	1	90		
21	20	M	90	0	90	0	72		

A screenshot of a Microsoft Excel spreadsheet titled "Association between numerical variables". The data is organized into columns A through I. Columns A, B, C, D, E, G, H, and I contain numerical values. Column F contains categorical values: 'F' for females and 'M' for males. Row 22 is labeled "MALE" and row 23 is labeled "FEMALE". Row 24 contains descriptive statistics: Sx = 9.57463755, yBar = 82.25, yBar = 78.66666667, RhoPS = 0.1881086147, and -0.1881086147. Row 25 contains values 69.125 and 89.83333333. Row 26 contains values -0.9635800872 and 0.935800872. Row 27 contains values 0.1881086147 and -0.1881086147.

A	B	C	D	E	F	G	H	I
8	7	F	91	1	91	1	91	
9	8	F	85	1	85	1	90	
10	9	F	69	1	69	1	90	
11	10	F	75	1	75	1	89	
12	11	M	92	0	92	0	68	
13	12	F	79	1	79	1	92	
14	13	M	71	0	71	0	72	
15	14	M	94	0	94	0	67	
16	15	F	86	1	86	1	92	
17	16	F	75	1	75	1	81	
18	17	F	90	1	90	1	93	
19	18	M	84	0	84	0	70	
20	19	F	91	1	91	1	90	
21	20	M	90	0	90	0	72	
22	MALE	p0	0.4	8	658	553		
23	FEMALE	p1	0.6	12	944	1078		
24		Sx	9.57463755		10.80192478			
25		yBar	82.25		69.125			
26		yBar	78.66666667		89.83333333			
27		RhoPS	0.1881086147		-0.9635800872			
28			-0.1881086147		0.935800872			

For this we have a measure which we refer to as a point bi-serial correlation coefficient. So, let X be a numerical variable and Y be a categorical variable. In our example, X is my marks it is the numerical variable, Y is the gender which is the categorical variable. I am assuming a dichotomous variable, gender has two levels, again I have a level of female and male. How do I compute the point bi-serial correlation between these two variables? The first thing which we do is the following.

We go back, you group the data into two sets based on the value of the dichotomous variable Y that is I am saying Y takes a value 0 or 1. Here in my example I have assumed Y takes the value if the gender is male and Y takes the 1 if the gender is female. And that is how I have coded my data. So, you can go back and see that this is how we have coded the data here. So, if I have Y , it takes the value 1 if it is a female and 0 if it is a male. Group the data into two sets based on the value. So, this data again I have, so this data I have 1, 71 this is the same data.

So, in this two sets I have two groups. What are the two groups? One is a female group and one is the male group. So, in the first step, in the second step you calculate the mean values of the two groups. How do I do this? One way to do this is, I can just find out what are the total number of females. So, if I count the total number of females it is 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 females. So out of 20, I have 12 females and 8 males. So, I have 8 males, so my 0, this is my male.

So, if I am summarizing it males, I have 8 males and I have 12 female students. So, the proportion is nothing but $\frac{8}{20}$ which is going to be 0.4 and $12/20$ which is 0.6. So, that is how and I am referring that to P_0 and P_1 . Now, what is the next step? The second step says, calculate the mean values of these two groups. So, how many females do I have? I have 12 females. What is the mean value?

It is going to be $71, 71 + 67 + 65 + \dots$, so forth I can keep doing $91 + 85 + 69 + 75 + 79 + 86 + 75 + 90 + 91$, which is 944. A simpler way to compute this is through using sum if function. How do you use a sum if function will be taught in the tutorial. It is a simpler (func) way that is if I have a female and just computing all those data points which correspond to the female, so I can use what is called the sum if function. So, this 944 represents the total marks obtained by female students.

Similarly, 658 is the total marks obtained by all men students or male students put together. Now, what is the mean of this group? I have 12 students in this group, so the mean of this group which is going to be nothing but $944/12$ which is equal to 78.66. So, \bar{y}_1 is giving me the mean of the female group. What is the average mark of the female students? Similarly the average mark of the male students is 82.25. So, if you go back to step two, it is saying that compute y_0 and \bar{y}_1 which are the mean values of the group.

So, y_0 here is the mean of the group of male students, \bar{y}_1 is the mean of the group of female students. Then I can find out what is the proportion of observation. Again I know that P_0 is going to be $8/20$, $p_1 = \frac{12}{20}$, this is the proportion of observations of the group which I am coding 0 and this is the group I am coding 1. Finally S_x, S_x, X is my numerical, it is a single numerical variable which is the marks. Remember when I have numerical variable, I can define what I call a standard deviation measure which is telling me about the variability of that particular variable.

So, S_x is the standard deviation of this variable. So, variable, entire variable which is $E2$ to $E21$. Then the point serial, point bi-serial correlation measure is defined as $y_0 - y_1 S_x$ divided by $P_0 P_1$. Certain books use n_0 and n_1, n_0 is the number of observations. So, p_0 is basically n_0 by total let, if n is the total number of

observation and n_0 is the number of observations in my 0th group, the group that is coded 0, n_1 is a number of observations in the group that is coded 1.

Then $n_0 \cdot p_0 = p_0 \cdot n_0$, p_1 is $n_1 \cdot p_1$. Some books or some authors instead of looking at this as $p_0 \times p_1$ use a sample correction which is (the) only this term is replaced by $n_0 \cdot n_1 \cdot n - 1 \times n_1 \cdot p_1$. So, you can see again this $n - 1$ which keeps playing a role always when you talk about sample characteristics. So, this is how you compute a point bi-serial correlation coefficient of a data which has both numerical and categorical variable.

(Refer Slide Time: 17:19)

A screenshot of an Excel spreadsheet titled "Association between numerical variables". The data consists of two main sections: a raw data section and a summary statistics section.

Raw Data Section:

	B	C	D	E	F	G	H	I
10	9	F	69	1	69	1	90	
11	10	F	75	1	75	1	89	
12	11	M	92	0	92	0	68	
13	12	F	79	1	79	1	92	
14	13	M	71	0	71	0	72	
15	14	M	94	0	94	0	67	
16	15	F	86	1	86	1	92	
17	16	F	75	1	75	1	81	
18	17	F	90	1	90	1	93	
19	18	M	84	0	84	0	70	
20	19	F	91	1	91	1	90	
21	20	M	90	0	90	0	72	
22	MALE	p0	0.4	8	658		553	
23	FEMALE	p1	0.6	12	944		1078	

Summary Statistics Section:

	Sx	9.57463755	10.80192478
24	y0bar	82.25	69.125
25	y1bar	78.66666667	89.83333333
26	RhoPS	0.1881086147	-0.9635800872
27		-0.1881086147	0.9635800872
28			
29			
30			

So, for this dataset again going back I have computed the point bi-serial coefficient which is .188 and this for the next data set is $-.96$. Again is there a difference between .188 and $-.188$. See, this 0 and 1 I could have change female to be 0 and male to be 1 then I would have got a different sign. That is what I have computed here, there is no, since this is, there is no order, there is no hard and fast rule that female should be 1 and male should be 0. I could have change the ordering.

(Refer Slide Time: 18:03)

The image shows two screenshots of Microsoft Excel spreadsheets. Both screenshots have a large circular watermark in the center with the text "INDIAN INSTITUTE OF STATISTICS" and "LOGY MADRAS".

Screenshot 1: This screenshot shows a table with columns A through I. Column A is labeled "Gender" and column B is labeled "Marks". Column C is labeled "Gender-coded" and column D is labeled "Marks". Column E is labeled "Gender-coded" and column F is labeled "Marks". Column G is labeled "Gender-coded" and column H is labeled "Marks". The formula in cell E2 is =if(B2="F",0,1). The formula in cell E3 is =if(B3="M",1,0). The formula in cell E4 is =if(B4="F",0,1). The formula in cell E5 is =if(B5="M",1,0). The formula in cell E6 is =if(B6="F",0,1). The formula in cell E7 is =if(B7="M",1,0). The formula in cell E8 is =if(B8="F",0,1). The formula in cell E9 is =if(B9="M",1,0). The formula in cell E10 is =if(B10="F",0,1). The formula in cell E11 is =if(B11="M",1,0). The formula in cell E12 is =if(B12="F",0,1). The formula in cell E13 is =if(B13="M",1,0). The formula in cell E14 is =if(B14="F",0,1). The formula in cell E15 is =if(B15="M",1,0). The formula in cell E16 is =if(B16="F",0,1). The formula in cell E17 is =if(B17="M",1,0). The formula in cell E18 is =if(B18="F",0,1). The formula in cell E19 is =if(B19="M",1,0). The formula in cell E20 is =if(B20="F",0,1). The formula in cell E21 is =if(B21="M",1,0).

	A	B	C	D	E	F	G	H	I
1		Gender	Marks	Gender-coded	Marks	Gender-coded	Marks		
2	1	F	71	=if(B2="F",0,1)	71	1	86		
3	2	F	67	=if(B3="M",1,0)	67	1	92	944	
4	3	F	65	EXAMPLE	65	1	92		
5	4	M	69	IF(A4="Yes","A2 Is True","A2 Is not True")	69	0	68		
6	5	M	75	Reduces one value if a logical expression is TRUE and another if it is FALSE.	75	0	70		
7	6	M	83	Logical_expression	83	0	66		
8	7	F	91	=IF(A8="Yes",A2 Is True,A2 Is not True)	91	1	91		
9	8	F	85	value_if_true	85	1	90		
10	9	F	69	value_if_false	69	1	90		
11	10	F	75	Learn more	75	1	89		
12	11	M	92		92	0	68		
13	12	F	79		79	1	92		
14	13	M	71		71	0	72		
15	14	M	94		94	0	67		
16	15	F	86		86	1	92		
17	16	F	75		75	1	81		
18	17	F	90		90	0	93		
19	18	M	84		84	1	70		
20	19	F	91		91	0	90		
21	20	M	90		90	1	72		
22	FEMALE	p0	0.6	12	944	0	89	1078	
23	MALE	p1	0.4	8	658	1	68	553	
24			Sx	9.57463755			10.80192478		
25			y0bar	78.66666667			89.83333333		
26			y1bar	82.25			69.125		
27			RhoPS	-0.1881086147			0.9635800872		
28				0.1881086147			-0.9635800872		
29									
30									
31									

Screenshot 2: This screenshot shows a table with columns A through I. Column A is labeled "Gender" and column B is labeled "Marks". Column C is labeled "Gender-coded" and column D is labeled "Marks". Column E is labeled "Gender-coded" and column F is labeled "Marks". Column G is labeled "Gender-coded" and column H is labeled "Marks". The formula in cell E2 is =(E25-E26)/E24)*sqrt((D22/20)*(D23/19)).

	A	B	C	D	E	F	G	H	I
11	10	F	75	0	75	0	89		
12	11	M	92	1	92	1	68		
13	12	F	79	0	79	0	92		
14	13	M	71	1	71	1	72		
15	14	M	94	1	94	1	67		
16	15	F	86	0	86	0	92		
17	16	F	75	0	75	0	81		
18	17	F	90	0	90	0	93		
19	18	M	84	1	84	1	70		
20	19	F	91	0	91	0	90		
21	20	M	90	1	90	1	72		
22	FEMALE	p0	0.6	12	944	0	89	1078	
23	MALE	p1	0.4	8	658	1	68	553	
24			Sx	9.57463755			10.80192478		
25			y0bar	78.66666667			89.83333333		
26			y1bar	82.25			69.125		
27			RhoPS	-0.1881086147			0.9635800872		
28				0.1881086147			-0.9635800872		
29									
30									
31									

So, I could have changed the ordering to be if it is female it is 0 and it is 1 here, otherwise. So, you can see that there is a flip in my this one, I have 12 males. So now this would become a female, because I had 12 females, sorry this would become a female now. This would be a male. 8 out of 12, this has, the standard deviation remains the same because I have not changed anything. My \bar{y}_0 now is going to be your $E22$ by $D22$, $y_{\bar{1}} \bar{1} \bar{2}$ is going to be $\frac{658}{8}$.

Now, you can see that the point bi-serial which was earlier .188 is now -.188. So depending on what is your coding, you are going to get this point bi-serial correlation coefficient that is the key difference between what I wanted to, what I wanted you to observe between numerical data, I have

a strong positive, or a strong negative, here it depends on how you are looking at the coding and that decides on what is your point bi-serial correlation coefficient.

So, whether it is a very, so here I can, so basically in absolute terms this would lie between 0 and 1 and it is closer to 0, it says that in this dataset my gender really has no association with the marks. And that is sort of validated by my scatter plot also for my first example which said that irrespective of which gender you are from, your marks seem to be distributed in the same way.

(Refer Slide Time: 20:15)

The image shows two screenshots of a software interface titled "Association between numerical variables".

Screenshot 1: A data table with columns labeled A through I. Columns B, C, E, and G are bolded. Row 1 contains column headers: Gender, Marks, Gender-coded, Marks, Gender-coded, Marks. Rows 2 through 21 contain data points. Row 21 is highlighted in yellow. Row 22 is a summary row for 'MALE' with values p0, 0.4, 8, 658, and 553. Row 23 is for 'FEMALE' with values p1, 0.6, 12, 944, and 1078. Row 24 is for statistics Sx, 9.57463755, and 10.80192478. Row 25 is for y0bar, 82.25, and 69.125. Row 26 is for y1bar, 78.66666667, and 89.83333333. Row 27 is for RhoPS, 0.1881086147, and -0.9635800872. Row 28 is for -0.1881086147, and 0.9635800872.

	A	B	C	D	E	F	G	H	I
1		Gender	Marks	Gender-coded	Marks	Gender-coded	Marks		
2	1	F	71	0	71	0	86		
3	2	F	67	0	67	0	92	944	
4	3	F	65	0	65	0	92		
5	4	M	69	1	69	1	68		
6	5	M	75	1	75	1	70		
7	6	M	83	1	83	1	66		
8	7	F	91	0	91	0	91		
9	8	F	85	0	85	0	90		
10	9	F	69	0	69	0	90		
11	10	F	75	0	75	0	89		
12	11	M	92	1	92	1	68		
13	12	F	79	0	79	0	92		
14	13	M	71	1	71	1	72		
15	14	M	94	1	94	1	67		
16	15	F	86	0	86	0	92		
17	16	F	75	0	75	0	81		
18	17	F	90	0	90	0	93		
19	18	M	84	1	84	1	70		
20	19	F	91	0	91	0	90		
21	20	M	90	1	90	1	72		
22	MALE	p0	0.4	8	658		553		
23	FEMALE	p1	0.6	12	944		1078		
24			Sx	9.57463755		10.80192478			
25			y0bar	82.25		69.125			
26			y1bar	78.66666667		89.83333333			
27			RhoPS	0.1881086147		-0.9635800872			
28				-0.1881086147		0.9635800872			

Screenshot 2: A data table with columns labeled A through I. Rows 9 through 21 contain data points. Row 22 is a summary row for 'MALE' with values p0, 0.4, 8, 658, and 553. Row 23 is for 'FEMALE' with values p1, 0.6, 12, 944, and 1078. Row 24 is for statistics Sx, 9.57463755, and 10.80192478. Row 25 is for y0bar, 82.25, and 69.125. Row 26 is for y1bar, 78.66666667, and 89.83333333. Row 27 is for RhoPS, 0.1881086147, and -0.9635800872. Row 28 is for -0.1881086147, and 0.9635800872.

	A	B	C	D	E	F	G	H	I
9	8	F	85	1	85	1	90		
10	9	F	69	1	69	1	90		
11	10	F	75	1	75	1	89		
12	11	M	92	0	92	0	68		
13	12	F	79	1	79	1	92		
14	13	M	71	0	71	0	72		
15	14	M	94	0	94	0	67		
16	15	F	86	1	86	1	92		
17	16	F	75	1	75	1	81		
18	17	F	90	1	90	1	93		
19	18	M	84	0	84	0	70		
20	19	F	91	1	91	1	90		
21	20	M	90	0	90	0	72		
22	MALE	p0	0.4	8	658		553		
23	FEMALE	p1	0.6	12	944		1078		
24			Sx	9.57463755		10.80192478			
25			y0bar	82.25		69.125			
26			y1bar	78.66666667		89.83333333			
27			RhoPS	0.1881086147		-0.9635800872			
28				-0.1881086147		0.9635800872			

Now, if we go to the second dataset, which is this F and G columns and do the same exercise for this dataset. Let me revert to my earlier coding of female is, does not make a difference but I just wanted to revert to my earlier coding and compute the point bi-serial. So, now again there are couple of things which I want you to see here. You can see that the total marks obtained by females here is way high than the total marks obtained by males, whereas here it was 944, 658, here it is much higher.

If you look at the group means, the means in the first example were close to each other, the total mean of the male group was about 82 and the female was about 78, in fact males had a higher mean than the females here. Whereas here you can see that the males are, the female mean is about close to 90 whereas the male mean is close to 70, there is a huge difference, here there is not too much of a difference.

The standard deviation is again not very different this is 9.5 and 10.8 they are not very different from each other, but when you look at the point bi-serial correlation measure, this is very much close to one, in fact it is .963. Again that is something which is validated here by saying that yes the gender seems to play a role when I am seeking the association between gender and marks in this example.

So, we stop here. What we have seen in this week is understand about association between variables. We looked at association between categorical variables, we looked at association between numerical variables and we looked at association between a numerical and a categorical variable. Here we just restricted our attention to a categorical variable which has only two categories and we introduced what is called the point bi-serial correlation coefficient.

Statistics for Data Science – 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Week - 4
Tutorial - 1

(Refer Slide Time: 0:14)

Statistics for Data Science - 1

Week 4 Tutorial Questions

Association between two variables

Syllabus covered:

- Use of two-way contingency tables to understand association between two categorical variables.
- Understand association between numerical variables through scatter plot; compute and interpret correlation.
- Understand relationship between a categorical and numerical variable.

Hello statistics students. In this tutorial, we are going to do questions based on week 4 topics which are these the contingency tables and association between variables. So, let us begin with our first question.

(Refer Slide Time: 0:29)

Q In a college, a charity organisation sought donations from various individuals and categorised them as Students, Staff, and Research Scholars. The following figure shows a contingency table for the groups of individuals and their attitudes towards the donation.

Group	Attitude to donation		Total	W.P	U.P
	Willing	Unwilling			
Students	835	65	900	75.6%	24.4%
Staff	77	23	100	77%	23%
Research scholar	110	90	200	55%	45%
Total	1022	178	1200	85.2%	14.8%

Based on the above table answer the questions 1 , 2 , 3 , 4 and 5.

Which group is more willing to donate?

Students

Staff

Research Scholars

All have an equal attitude

What is the percentage (%) contribution of Students in the total positive attitude towards donation?

75.6

So, this is our first question; in a college, a charity organisation sought donations from various individuals and categorise them as students, staff and research scholars. The

following figure shows the contingency table for the groups of individuals and their attitudes towards the donations. So, there are two kinds of attitudes we are looking at, one is willing, the other is unwilling and we have the respective numbers for these. And they are asking which group is more willing to donate?

So, I think we should be comparing the percentages here. So, let me add two more columns, willing percentage and unwilling percentage. So, WP, the willing percentage in students is going to be $\frac{835}{900} \times 100$. So, that is essentially $835/9$ which should come to about 92.8%. And that would give us the, remaining unwilling percentage would be the remaining 7.2 percent. And in the next case we have 77 and 23 coming to a total of 100. So, this is easy.

So, the staff is 77% inclined and 23% not interested and then this is 110/200 which just gives us 55% willing and remaining 45% unwilling. So, if you looked at total as well, the total would be 85.2 percent and so the remaining which is 14.8 percent are unwilling. Now, the question is which group is more willing to donate and it is pretty clear that 92.8% is the maximum in the willing percentages, so our answer should be students. Students are most willing to donate in this case.

(Refer Slide Time: 3:11)

Group		Willing	Unwilling	Total	WP	UP
		Students	835	65	900	72%
	Staff	77	23	100	77%	23%
	Research scholar	110	90	200	55%	45%
	Total	1022	178	1200	85.2%	14.8%

$\frac{835}{900} \times 100$

Based on the above table answer the questions 1, 2, 3, 4 and 5.

Which group is more willing to donate?

- Students
- Staff
- Research Scholars
- All have an equal attitude

What is the percentage (%) contribution of Students in the total positive attitude towards donation?

- 75.6
- 81%
- 86.78
- 67.34

$\frac{835}{1022} \times 100 \approx 81.7\%$

What is the percentage of students who are unwilling to donate?

- 7.111



Then we have what is the percentage contribution of students in the total positive attitude towards donation? Now this is where we have to be a little bit careful in what we are doing. So, the percentage contribution of students in the total positive attitude is not this 92.8%, it is in fact, 835 as a percentage of 1022. So, of the total willing candidates, how much percent is being contributed by students? So, that would be $\frac{835}{1022} \times 100$ and this comes to about roughly

81.7%. So, they are asking the percentage contribution, which means our answer is 81.7. Moving further.

(Refer Slide Time: 4:15)

- Students
- Staff
- Research Scholars
- All have an equal attitude

What is the percentage (%) contribution of Students in the total positive attitude towards donation?

- 75.6
- 71.1
- 86.78
- 67.34

$\frac{835}{1022} \times 100 \approx 81.7\%$

What is the percentage of students who are unwilling to donate?

- 7.11
- 92.88
- 1.54
- None of the above

↳



In a college, a charity organisation sought donations from various individuals and categorised them as Students, Staff, and Research Scholars. The following figure shows a contingency table for the groups of individuals and their attitudes towards the donation.

Group	Attitude to donation			W.P	U.P
	Students	Willing	Unwilling		
Students	835	65	900	72%	28%
Staff	77	23	100	77%	23%
Research scholar	110	90	200	55%	45%
Total	1022	178	1200	85.2%	14.8%

Based on the above table answer the questions 1, 2, 3, 4 and 5.

Which group is more willing to donate?

- Students
- Staff
- Research Scholars
- All have an equal attitude

What is the percentage (%) contribution of Students in the total positive attitude towards donation?

- 75.6



What is the percentage of students who are unwilling to donate? This we had already calculated. 7.2 we had taken the approximation there it is actually supposed to be I believe 7.11. So, that should be our closest answer.

(Refer Slide Time: 4:34)



Suppose the means of the donations (in rupees) from these groups (counting only those donated) are as follows:

$$\begin{array}{l} \text{Student} \quad 100 \times 835 = ₹ 83,500 \\ \text{Staff} \quad 340 \times 77 = ₹ 26,180 \\ \text{Research scholar} \quad 240 \times 110 = ₹ 26,400 \end{array}$$

Which group donated more money towards charity?

- Student
- Staff
- Research Scholar
- Student and Staff
- Staff and Research scholar
- All donate the same amount

Which group donated more money towards charity per head on average?

- Student
- Staff
- Research scholar



In a college, a charity organisation sought donations from various individuals and categorised them as Students, Staff, and Research Scholars. The following figure shows a contingency table for the groups of individuals and their attitudes towards the donation.

Group	Attitude to donation		Total	Willing	Unwilling	
	Students	Staff		72%	28%	
Students	835	65	900	791	231	
Staff	77	23	100	55	45	
Research scholar	110	90	200	154	46	
Total	1022	178	1200	852	148	

Based on the above table answer the questions 1 , 2, 3 , 4 and 5.

Which group is more willing to donate?

- Students
- Staff
- Research Scholars
- All have an equal attitude

What is the percentage (%) contribution of Students in the total positive attitude towards donation?

- 75.6

Now we going to slightly more interesting question where the means of the donations from these groups are also given. So, the means are given. So, student in average has given 100, staffs has on average given 340 per staff member and research scholars on average have given 240 rupees per research scholars. So, although students are more inclined to donate, they are likely to donate lesser money as we can observe here. Now, which group donated more money towards the charity?

So, we now need the actual donations which supposed to come from these values. The number of people who have been willing to donate. So, 835, 77 and 110. That is, 835, 77 and 110. So, what is the total amount of money donated by students? That would be the average, the mean of their donations multiplied by the number of students that donated. So, that is

going to give us rupees 83500, whereas staff is 340 per head multiplied by 77 staff members, so that is rupees 26180 and lastly we have research scholars who gave 26400 rupees.

So, as you can see although the staff members actually had a high percentage of donation, being willing to donate and actually the largest average donation per head. Simply because their numbers are really small, their total contribution is the least in the lot. Research scholars have actually contributed about 220 more than staff on the whole. So, which group donated more money towards charity? Clearly, students.

(Refer Slide Time: 7:04)


 IIT Madras
ONLINE DEGREE

<p>Staff $340 \times 77 = ₹ 26,180$</p> <p>Research scholar $240 \times 110 = ₹ 26,400$</p> <p>Which group donated more money towards charity?</p> <ul style="list-style-type: none"> <input checked="" type="radio"/> Student <input type="radio"/> Staff <input type="radio"/> Research Scholar <input type="radio"/> Student and Staff <input type="radio"/> Staff and Research scholar <input type="radio"/> All donate the same amount <p>Which group donated more money towards charity per head on average?</p> <ul style="list-style-type: none"> <input type="radio"/> Student <input checked="" type="radio"/> Staff <input type="radio"/> Research scholar <input type="radio"/> Student and Staff <input type="radio"/> Staff and Research scholar <input type="radio"/> All donate the same amount 	<p>Suppose the means of the donations (in rupees) from these groups (counting only those donated) are as follows:</p> <table border="0" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 30%; vertical-align: top;"> <p>Student $100 \times 835 = ₹ 83,500$</p> </td> <td style="width: 10%;"></td> <td style="width: 60%; vertical-align: top;"> <p>Staff $340 \times 77 = ₹ 26,180$</p> </td> </tr> <tr> <td style="width: 30%; vertical-align: top;"> <p>Research scholar $240 \times 110 = ₹ 26,400$</p> </td> <td style="width: 10%;"></td> <td style="width: 60%; vertical-align: top;"></td> </tr> </table> <p>Which group donated more money towards charity?</p> <ul style="list-style-type: none"> <input checked="" type="radio"/> Student <input type="radio"/> Staff <input type="radio"/> Research Scholar <input type="radio"/> Student and Staff <input type="radio"/> Staff and Research scholar <input type="radio"/> All donate the same amount <p>Which group donated more money towards charity per head on average?</p> <ul style="list-style-type: none"> <input type="radio"/> Student <input checked="" type="radio"/> Staff <input type="radio"/> Research scholar 	<p>Student $100 \times 835 = ₹ 83,500$</p>		<p>Staff $340 \times 77 = ₹ 26,180$</p>	<p>Research scholar $240 \times 110 = ₹ 26,400$</p>			<p>Staff $340 \times 77 = ₹ 26,180$</p> <p>Research scholar $240 \times 110 = ₹ 26,400$</p> <p>Which group donated more money towards charity?</p> <ul style="list-style-type: none"> <input checked="" type="radio"/> Student <input type="radio"/> Staff <input type="radio"/> Research Scholar <input type="radio"/> Student and Staff <input type="radio"/> Staff and Research scholar <input type="radio"/> All donate the same amount <p>Which group donated more money towards charity per head on average?</p> <ul style="list-style-type: none"> <input type="radio"/> Student <input checked="" type="radio"/> Staff <input type="radio"/> Research scholar
<p>Student $100 \times 835 = ₹ 83,500$</p>		<p>Staff $340 \times 77 = ₹ 26,180$</p>						
<p>Research scholar $240 \times 110 = ₹ 26,400$</p>								

And then the next question is which group donated more money towards charity per head on average? So, that would be the staff. As we had seen 340 is the max that has shown up in the means.

Statistics for Data Science - 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Week - 4
Tutorial - 2

(Refer Slide Time: 0:14)

The following data displays the temperature of the day (K) and the number of people at the beach.

Temperature	No.of people
293	1500
295	1300
299	1150
300	800
305	500
297	1200
308	456
312	200

Determine the sample correlation coefficient between the number of people at the beach and the temperature of the day and interpret the result.

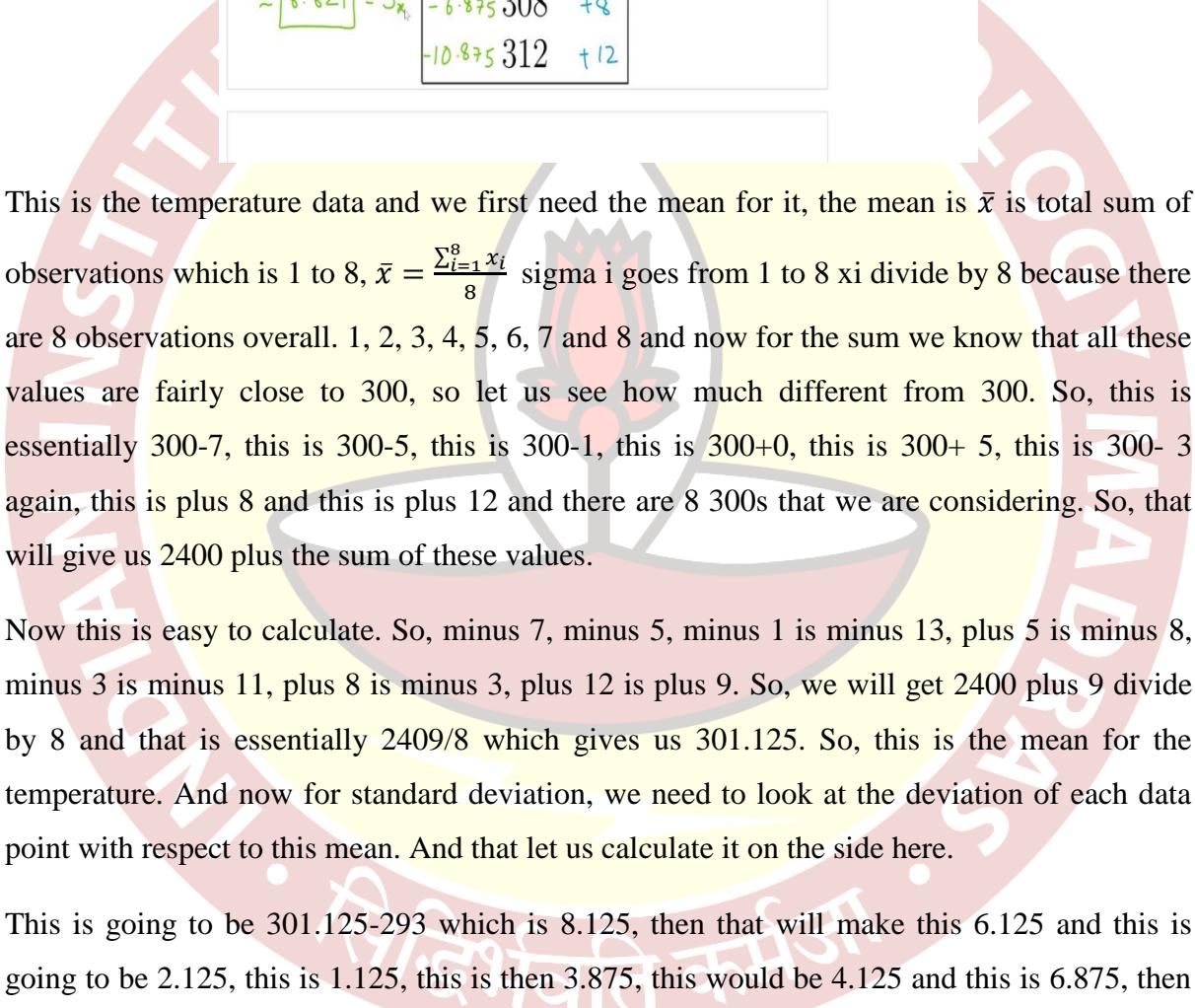
Sample Correlation Coefficient =
$$\frac{\text{Sample Covariance}}{\text{Sample Std. deviations.}}$$



For our second question, we are given the data regarding the temperature of the day and the number of people at the beach. So, here this is the temperature of the day in Kelvin, 293 kelvin, 295 kelvin and so on. And this is the number of people who visited the beach on that day. Now we are supposed to determine the sample correlation coefficient between these two variables and at the end of it we are supposed to interpret the result.

And let us see how this is supposed to be done. We want the sample correlation coefficient, which you know the formula is $\frac{xy}{\sqrt{x^2 - \bar{x}^2} \sqrt{y^2 - \bar{y}^2}}$. and here what we mean by these terms, this (S_{xy}) is the sample covariance between these two variables, whereas S_x and S_y are independently the sample standard deviations of these two variables. So, this means that we have to calculate the sample covariance and calculate the sample standard deviations separately. So, let us begin with the standard deviations.

(Refer Slide Time: 1:59)



Temperature		
8.125	293	-7
6.125	295	-5
2.125	299	-1
1.125	300	0
-3.875	305	+5
4.125	297	-3
-6.875	308	+8
-10.875	312	+12

IIT Madras
ONLINE DEGREE

$$\bar{x} = \frac{\sum_{i=1}^8 x_i}{8}$$

$$= \frac{2400 + 9}{8}$$

$$= \frac{2409}{8}$$

$$= 301.125$$

$$\text{Standard Deviation } S_x = \sqrt{\frac{\sum_{i=1}^8 (\bar{x} - x_i)^2}{8}}$$

$$= \sqrt{\frac{306.875}{8}}$$

$$= \sqrt{38.359375}$$

$$\approx 6.621 = S_x$$

This is the temperature data and we first need the mean for it, the mean is \bar{x} is total sum of observations which is 1 to 8, $\bar{x} = \frac{\sum_{i=1}^8 x_i}{8}$ sigma i goes from 1 to 8 x_i divide by 8 because there are 8 observations overall. 1, 2, 3, 4, 5, 6, 7 and 8 and now for the sum we know that all these values are fairly close to 300, so let us see how much different from 300. So, this is essentially 300-7, this is 300-5, this is 300-1, this is 300+0, this is 300+ 5, this is 300- 3 again, this is plus 8 and this is plus 12 and there are 8 300s that we are considering. So, that will give us 2400 plus the sum of these values.

Now this is easy to calculate. So, minus 7, minus 5, minus 1 is minus 13, plus 5 is minus 8, minus 3 is minus 11, plus 8 is minus 3, plus 12 is plus 9. So, we will get 2400 plus 9 divide by 8 and that is essentially 2409/8 which gives us 301.125. So, this is the mean for the temperature. And now for standard deviation, we need to look at the deviation of each data point with respect to this mean. And that let us calculate it on the side here.

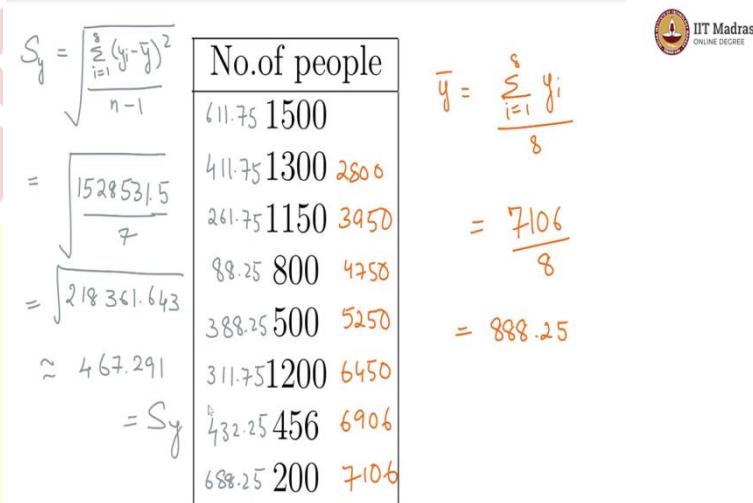
This is going to be 301.125-293 which is 8.125, then that will make this 6.125 and this is going to be 2.125, this is 1.125, this is then 3.875, this would be 4.125 and this is 6.875, then that would be this would be 10.875. Obviously, I am ignoring the signs if we are particular about the signs, then \bar{x} minus x_i , this will be positive, positive, positive, positive, this is negative, this is again positive, this is negative and this is negative.

But we are not interested in the signs because we are going to square them and then do the division. So, for sample standard deviations I am going to do the sum of all these deviations

square, $s = \sqrt{\frac{\sum_{i=1}^8 (x_i - \bar{x})^2}{n-1}}$. This requires considerable calculation. I will not show that in the tutorial but if you have calculators you can do it and verify for yourself.

What we get at the end of the calculation is $\sqrt{\frac{306.875}{7}} = \sqrt{43.839} = 6.621$. So, this is our sample standard deviation for the x variable for the temperature variable. So, this is S_x . Now, let us calculate S_y .

(Refer Slide Time: 6:09)



Now for the y variables again we have to calculate the mean, the $\bar{y} = \frac{\sum_{i=1}^8 y_i}{8}$. So, sum would be giving us 1500 plus 1300 is 2800 plus 1150 would be 3950 plus 800 would give us 4750 plus 500 would give us 5250 plus 1200 would give us 6450 plus 456 would give us 6906 plus 200 gives us 7106. So, this is equal to 7106/8 which comes out to be 888.25.

Now once again let us calculate the deviations of each data point from the mean. I am again using the absolute values only. So, this value, 1500 minus 888.25 will give us 611.75. Then this next one will give us 200 lesser, 411.75 and the next one will give us 150 further lesser which is 261.75. And this one is 88.25 and this one is going to be 388.25, then this one is going to be 311.75 and this is 432.25 and the last one is 688.25.

So, we have pretty large numbers, again we are going to need a calculator for this. Anyway we are calculating the $s_y = \sqrt{\frac{\sum_{i=1}^8 (y_i - \bar{y})^2}{n-1}}$. So, that sum of the squares if we calculate it, we get

1528531.5/7. And this is equal to square root of 218361.643 which comes out to be roughly 467.291. So, this is our $S_y = 467/291$.

(Refer Slide Time: 9:17)



Temperature	No.of people
8.125 293	611.75 1500
6.125 295	411.75 1300
2.125 299	261.75 1150 $\bar{y} =$
1.125 300	88.25 800 888.25
3.875 305	388.25 500
4.125 297	311.75 1200
6.875 308	432.25 456
10.875 312	688.25 200

Now for the sample covariance we are going to use these deviations and the sign matters now. What is the sample covariance formula?

(Refer Slide Time: 9:29)



$x =$	$y =$
1.125 300	88.25 800 888.25
3.875 305	388.25 500
4.125 297	311.75 1200
6.875 308	432.25 456
10.875 312	688.25 200

$$S_{xy} = \frac{\sum_{i=1}^8 (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

The standard covariance formula is S_{xy} which is the sample covariance $s_{xy} = \frac{\sum_{i=1}^8 (x_i - \bar{x})(y_i - \bar{y})}{n-1=7}$.

So, every respective deviation the whole divided by n minus 1 which is 7.

(Refer Slide Time: 9:57)

Temperature	No.of people
-8.125	293
-6.125	295
-2.125	299
-1.125	300
+3.875	305
-4.125	297
+6.875	308
+10.875	312
$\bar{x} = 88.25$	
$\bar{y} = 888.25$	
+611.75	1500
+411.75	1300
+261.75	1150
-88.25	800
-388.25	500
+311.75	1200
-432.25	456
-688.25	200

So, the signs are of importance here, $x_i - \bar{x}$ is what we are looking at, here x_i is lesser than \bar{x} so this is negative, this is also negative, this is also negative, this is also negative, this is positive, this is again negative, this is positive and this is positive. So, we have 3 positive and 5 negative. Whereas over here, 888.25, 888.25 is the mean. So, 1500 is definitely greater, this is positive. This is also positive, this is also positive, this would be negative, this is also negative, this is positive, then this is also negative, this is also negative. Here, we have 4 positive and 4 negative.

(Refer Slide Time: 10:52)

+6.875	308	-432.25	456
+10.875	312	-688.25	200

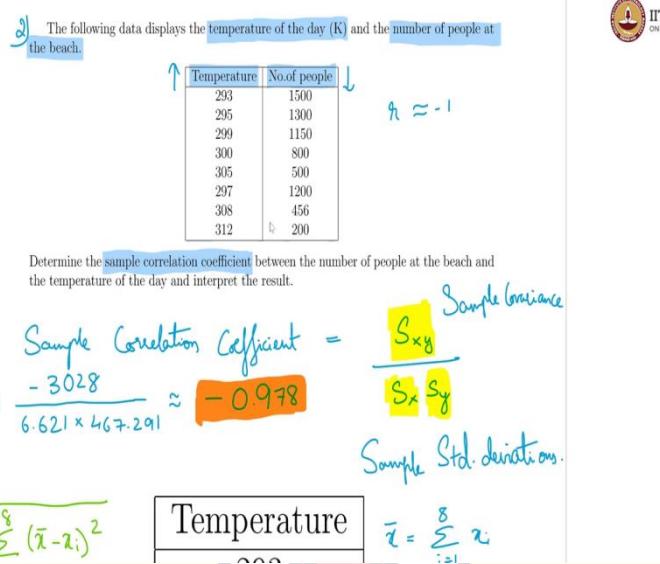
$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$\begin{aligned}
 &= \frac{1}{7} [(-8.125)(611.75) + (-6.125)(411.75) + (-2.125)(261.75) + \\
 &\quad (-1.125)(-88.25) + (3.875)(388.25) + (-4.125)(311.75) + \\
 &\quad (6.875)(-432.25) + (10.875)(-688.25)] \\
 &= \frac{-21196.25}{7} = -3028.0357 \\
 &\approx -3028 = S_{xy}
 \end{aligned}$$

Now we need to calculate these products respectively. If we put this down, this is what it will look like fairly elaborate and long calculation is required. We have minus 8.125 into 611.75

plus this whole thing, the whole entire sum multiplied by 1 by 7. The calculation gives us $\frac{-21196.25}{7}$, which gives us -3028.03571. So, roughly -3028 and this is what we are going to take as our S_{xy} . So, now we have these values, we have the sample covariance as -3028.

(Refer Slide Time: 12:04)



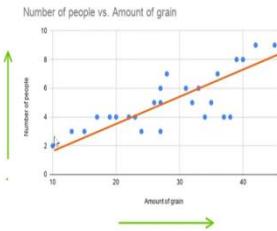
So, sample standard deviation of the y variable, the number of people is 467.291 and lastly the sample standard deviation of the x variable which is 6.621. And thus, our sample correlation coefficient would come out to be minus 3028 divided by 6.621 into 467.291 $\frac{-3028}{6.621 \times 467.291}$ which then comes out to be roughly -0.978. This is a value which is very, very close to -1.

So, we are saying our sample correlation coefficient r is almost -1 which indicates a very strong linear relationship but while 1 increases, the other decreases. And this is quite understandable. As the temperature increases, number of people going to the beach will decrease. So, that is the interpretation of our result. There is a strong linear relationship, strong correlation between these two variables and while one increases, the other decreases.

Statistics for Data Science - 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Week - 4
Tutorial - 3

(Refer Slide Time: 0:14)

 The following scatter plot shows the amount of grain consumed (in kg) versus the number of people in the household:



a) Then the association between the two variables is

- positive
- negative
- strong
- weak
- zero

Now, this is our third question, the following scatter plot shows the amount of grain consumed in kilogram versus the number of people in the household. So, this is the scatter plot. Then the association between the two variables is positive or negative, strong or weak or 0. So, first of all as the amount of the grain is increasing, we see that the number of people is also increasing.

There is that positive association between them so definitely positive is true, it is not negative. Is it strong or weak and this is also not 0 because definitely positive. Now in terms of strong or weak, I think we can do this by trying to see what a line fit will look like. It is quite clear that there seems to be a reasonably good line fit over here and this would indicate a fairly strong association between the variables.

It is not weak definitely because as the amount of grain is increasing, the number of people is also increasing. It should actually be the other way around, as number of people are increasing, the amount of grain consumed will also increase. So, this is a fairly strong linear relationship that we can observe here.

(Refer Slide Time: 1:47)

b) The sample standard deviation of number of people in a household is 1.93, and the sample standard deviation of the amount of the grain consumed is 9.36. If their sample covariance is 13.96, what is the sample correlation coefficient between the number of people in the household and the amount of grain consumed?

$$\frac{s_{xy}}{s_x s_y} = \frac{13.96}{1.93 \times 9.36}$$
$$r \approx 0.7728 > 0.5$$

In the second part of the question we are being asked, what is the sample correlation coefficient, that is what is being asked of us and what is given is the sample standard deviation of number of people which is 1.93 and sample standard deviation of the amount of grain which is 9.36 and lastly the sample covariance is also given to us which is 13.96. We have already seen in the previous problem that $r = \frac{s_{xy}}{s_x s_y}$.

And that gives us 13.96 divided by 1.93 into 9.36, and this is roughly 0.7728. So, our sample correlation coefficient is roughly 0.7728 > 0.5 and it is positive. So, our earlier detection that it is a strong and positive association is confirmed here because we are getting a fairly strong sample correlation coefficient.

Statistics for Data Science - 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Week - 4
Tutorial - 4

(Refer Slide Time: 0:14)

The following data gives the monthly income (in thousand of rupees) of female and male nurses of a hospital. Which gender can be said to be earning more?



Gender	Monthly Income
F	46
F	47
F	40
M	34
M	18
M	22
F	45
F	50
F	55
F	60
F	69
M	34
M	36
M	35
F	70
F	75
F	80
M	28
M	44
M	33

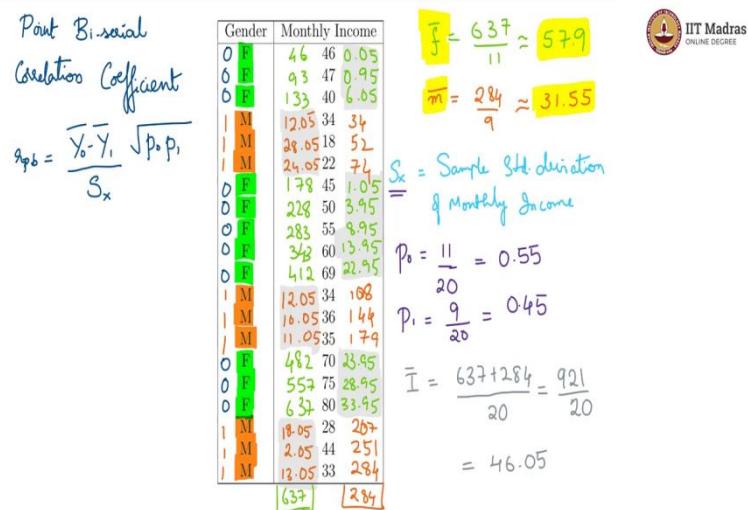
1) Find the gender with the greater mean of monthly income.

2) Determine whether the correlation between the variables is high or not.

Now for our fourth question, we have the monthly income in 1000s of rupees of female and male nurses. So, the gender is also given and the monthly income of that nurse is given. Which gender can be said to be earning more? So, which gender can be said to be earning more? So, this is layered question, one is which gender can be said to be earning more and the second part of that question is can we be sure that there is actually a relation between gender and monthly income?

So the part 1 which gender can be obtained from identifying the mean of the income for the females and the mean of income for males and then comparing the mean. But, the second part where we have to establish whether gender and monthly income actually have any connection which is basically the correlation, will tell us whether this difference in means is significant.

(Refer Slide Time: 1:39)



So, for this we are now going to have to find out the mean of the female nurses salaries and the mean of the male nurses salary. So, let us begin with female. So, let us identifying our female data points here, we have 3 and 4 and 5, this is 6, this 7, there is 8, this is 9, this is 10 and this is 11. So, we are saying there are 11 female nurses and let us calculate their sum. So, this would be 46, we are starting with 46 and then 47 will give us 93 and 40 will give us a 133.

And with 45 we get 178, then this gives us 228, this gives us 283, then we have 343, then we have 412, with 70 we get 482 and with this we get 557, then another 80 we get 637. So, so the female sum is 637 and there are 11 female nurses. Therefore, the mean of female monthly incomes is let us call it \bar{f} equal to 637 divided by 11 which is roughly equal to 57.9. Now in the case of men, this is male, this is male, this is male, again male, male, male and lastly here another 3.

So, we have 9 males overall. Let us note down that here, there are males overall. And for their sum again let us do the counting on this side. So, this 34, this would be 52, this gives us 74, moving on here we have 108 and 144 and then this gives us 179, there along with this 28 we get 207 plus 44 is 251 and lastly with 33 we get 284. And so the sum here for the male nurses is 284 and we know there are 9 of them, so let this be called \bar{m} , the mean for male nurses this is 284/9 which is roughly again 31.55.

So, evidently the female mean is better and now we come to the second part of the question, can we make such an inference at all? Is it okay for us to say that this gender is earning more,

is there that sort of correlation? And that we can only find out by finding the point bi-serial correlation coefficient. The r_{pb} that is the point bi-serial correlation coefficient is $\frac{(Y_0 - Y_1)\sqrt{P_0 P_1}}{S_x}$.

Now, what do these respective terms mean?

For that, we need to first need to first talk about the encoding of our gender variable as 0 and 1. So, there are two kinds of genders here, so we assign 0 to one and 1 to the other. Let us give 0 here to the female gender and let us make 1 for the male gender. So, this is a encoding that we are doing. And \bar{Y}_0 would in that case be \bar{f} which is the mean for the 0 variable and \bar{Y}_1 bar would be the mean for the one variable which is \bar{m} that is our 31.55.

And what is S_x ? S_x is the sample standard deviation of the monthly income variable. So, all these values here that we see in monthly income, they have their own sample standard deviation and that is the value that we give to S_x . Going further, P_0 is the proportion for the 0 variable which is number of 0s by the total number of values here. The number of 0s here 3 and 5 and 3 that gives us 11. So, we have 11 divided by how many are the total, there is an additional 3, 3 and 3 which gives 9.

So, $11/20$ the total is 20 so this is 0.55. And that would indicate P_1 is equal to 9 by 20 which gives 0.45. So, we have the values of \bar{Y}_0 and \bar{Y}_1 and P_0 and P_1 . We are yet to find the sample standard deviation of the monthly income. So, let us do that. We need to first find the mean for the monthly income which we can obtain by, I will call that value \bar{I} . The mean for monthly income, it should be the sum of these two, that is the total sum is the sum of these two.

So, 637 plus 284 divided by the total number of observations is 20 so we get 921/20. And that is essentially 46.05. Now what we do is, we write down the deviations of these values from the mean, from \bar{I} . So, 46 is 0.05 away. I am ignoring the signs because we are anyway going to square them and 47 is 0.95 away, 40 is 6.05 away, then 34 is 12.05 away, 18 is 28.05 away, 22 is 24.05 away, 45 is 1.05 away, 50 is 3.95 away.

55 is 8.95 away, 60 is 13.95 away, 69 is 22.95 away, 34 is 12.05 away, 36 is 10.05 away, 35 will be 11.05 away. And 70 is 23.95 away, 75 is 28.95 away, 80 is 33.95 away, 28 is 18.05 away, 44 is 2.05 away, 33 is 13.05 away. So, these are all our deviations and we would like to square them all, add them, divide them by n which in this case is 20 and perform the square root on that value in order to get the standard deviation.

(Refer Slide Time: 10:47)

0	F	283	55	9.95
0	F	343	60	13.95
0	F	412	69	21.95
1	M	12.05	34	1.08
1	M	10.05	36	1.44
1	M	11.05	35	1.79
0	F	482	70	33.95
0	F	552	75	28.95
0	F	632	80	33.95
1	M	18.05	28	2.04
1	M	2.05	44	2.51
1	M	12.05	33	2.84
		637	284	

$$P_0 = \frac{11}{20} = 0.55$$

$$P_1 = \frac{9}{20} = 0.45$$

$$\bar{x} = \frac{637+284}{20} = \frac{921}{20}$$

$$= 46.05$$



$$\sqrt{\frac{\sum_{i=1}^{20} (x_i - \bar{x})^2}{19}} = S_x = \sqrt{\frac{5798.95}{19}}$$

$$\approx \sqrt{305.2} \approx 17.47$$

nurses of a hospital. Which gender can be said to be earning more?

Point Biserial

Correlation Coefficient

$$r_{pb} = \frac{\bar{Y}_0 - \bar{Y}_1}{S_x} \sqrt{P_0 P_1}$$

$$= \frac{(57.9 - 31.55) \sqrt{0.55 \times 0.45}}{17.47}$$

$$= \frac{26.35 \times 0.4975}{17.47}$$

$$= \frac{13.109125}{17.47} \approx 0.75$$

Gender	Monthly Income			
0	F	46	46	0.05
0	F	93	47	0.95
0	F	133	40	6.05
1	M	12.05	34	1.08
1	M	28.05	18	5.2
1	M	24.05	22	7.4
0	F	179	45	1.05
0	F	228	50	3.95
0	F	283	55	9.95
0	F	343	60	13.95
0	F	412	69	21.95
1	M	12.05	34	1.08
1	M	10.05	36	1.44
1	M	11.05	35	1.79
0	F	482	70	33.95
0	F	552	75	28.95
0	F	632	80	33.95
1	M	18.05	28	2.04
1	M	2.05	44	2.51
1	M	12.05	33	2.84
		637	284	

$$\bar{Y}_0 = \frac{637}{11} \approx 57.9$$

$$\bar{Y}_1 = \frac{284}{9} \approx 31.55$$

S_x = Sample Std. deviation
of Monthly Income

$$P_0 = \frac{11}{20} = 0.55$$

$$P_1 = \frac{9}{20} = 0.45$$

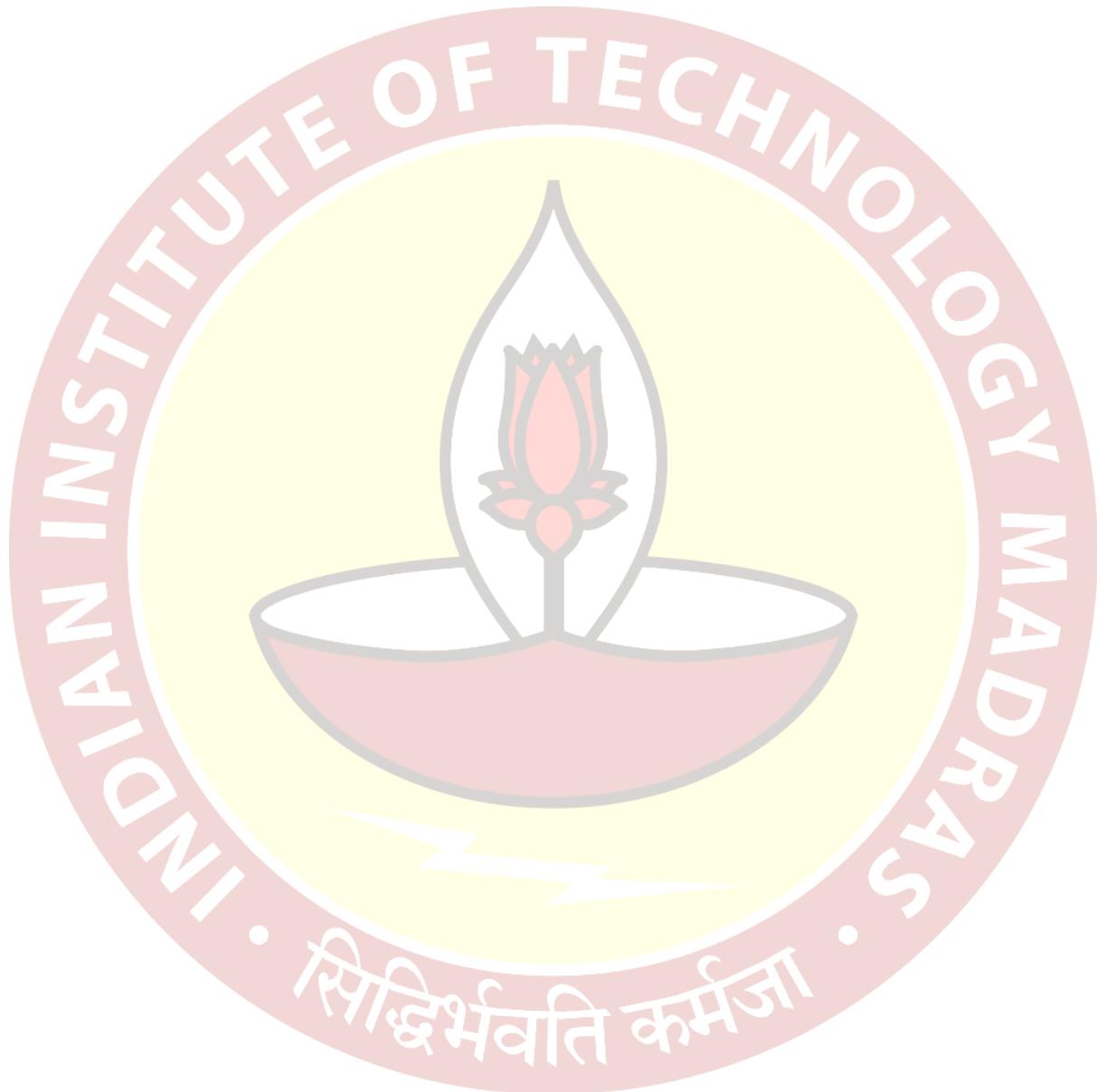
$$\bar{x} = \frac{637+284}{20} = \frac{921}{20}$$

$$= 46.05$$

Which is this calculation here $\sigma_x = \sqrt{\frac{\sum_{i=1}^{20} (x_i - \bar{x})^2}{n}}$. Now I will not do this calculation of this numerator here, what we do get for this is 5798.95, 5798.95/20 that gives us $\sqrt{289.9475}$ roughly which is again roughly 17.02. So, we finally have the value of s_x as well, we have P_1 , we have P_0 and \bar{Y}_0 and \bar{Y}_1 anyway.

So, \bar{Y}_0 and \bar{Y}_1 are also there. So, if we substitute all these values, $\bar{Y}_0 - \bar{Y}_1$ will give us 57.9 - 31.55, this whole thing multiplied by the $\sqrt{0.55 \times 0.45}$ whole divided by 17.02. 26.35 into 0.4975 which is the square root of 0.55 into 0.45 divided by 17.47 $r_{pb} = \frac{(Y_0 - Y_1)\sqrt{P_0 P_1}}{S_x} = \frac{(57.9 - 31.55)\sqrt{0.45 \times 0.55}}{17.02} = \frac{(26.35) \times 0.4975}{17.02} = \frac{13.109125}{17.02}$ which is coming out to be very close to 0.77.

And 0.77 is a fairly large correlation coefficient, it is very close to 1 and it is positive. So, there is a strong positive correlation between gender and monthly income. In that case, we can safely say that the female nurses earn better than male nurses in that hospital. (Thank).



**THIS BOOK IS
NOT FOR SALE
NOR COMMERCIAL USE**



Scan QR code to chat
with our WhatsApp bot

+91 9840776800



7850-999966
(Mon-Fri 9am-6pm)



Online Degree Office

3rd Floor, ICSR Building, IIT Madras, Chennai - 600036



onlinedegree.iitm.ac.in



support@onlinedegree.iitm.ac.in