Lecture Notes Unit I

Subject: Fundamentals of Data Science & Analytics

1. What is Data Science?

Data science involves using methods to analyze massive amounts of data and extract the knowledge it contains. Data science and big data evolved from statistics and traditional data management but are now considered to be distinct disciplines.

Data Science: Data Science is a field or domain which includes and involves working with a huge amount of data and uses it for building predictive, prescriptive and prescriptive analytical models. It’s about digging, capturing, (building the model) analyzing (validating the model) and utilizing the data (deploying the best model).

It is an intersection of Data and computing. It is a blend of the field of Computer Science, Business Management and Statistics together.

Big Data: It is huge, large or voluminous data, information or the relevant statistics acquired by the large organizations and ventures. Many software and data storage created and prepared as it is difficult to compute the big data manually.

It is used to discover patterns and trends and make decisions related to human behavior and interaction technology.

Example Applications:

Fraud and Risk Detection

Healthcare

Internet Search

Targeted Advertising

Website Recommendations

Advanced Image Recognition

Speech Recognition

Airline Route Planning

Gaming

Augmented Reality

2. What are all the difference between the Big Data and Data Science?

[redacted]	
Data Science	Big Data
[redacted]	

Data Science is an area.	Big Data is a technique to collect, maintain and process the huge information.
[redacted]	[redacted]
It is about collection, processing, analyzing and utilizing of data into various operations. It is more conceptual.	It is about extracting the vital and valuable information from huge amount of the data.
[redacted]	[redacted]
It is a field of study just like the Computer Science, Applied Statistics or Applied Mathematics, Data Base Management System.	It is a technique of tracking and discovering of trends of complex data sets.
[redacted]	[redacted]
The goal is to build data-dominant products for a venture.	The goal is to make data more vital and usable i.e. by extracting only important information from the huge data within existing traditional aspects.
[redacted]	[redacted]
Tools mainly used in Data Science includes SAS, R, Python, etc	Tools mostly used in Big Data includes Hadoop, Spark, Flink, etc.
[redacted]	[redacted]
It is a sub set of Data Science as mining activities which is in a pipeline of the Data science.	It is a super set of Big Data as data science consists of Data scrapping, cleaning, visualization, statistics and many more techniques.
[redacted]	[redacted]
It is mainly used for scientific purposes.	It is mainly used for business purposes and customer satisfaction.

[redacted]	[redacted]
Uses mathematics and statistics extensively along with programming skills to develop a model to test the hypothesis and make decisions in the business	Used by businesses to track their presence in the market which helps them develop agility and gain a competitive advantage over others.
[redacted]	[redacted]
Internet search, digital advertisements, text-to-speech recognition, risk detection, and other activities.	Telecommunication, financial service, health and sports, research and development, and security and law enforcement

3. What are all the characteristics of big data?

The characteristics of big data are explained with 'Five V' approach. If it satisfy the five characteristics then it is known as Big Data.

■ Volume--how much data is there? To determine the value of data, size of data plays a very crucial role. If the volume of data is very large then it is actually considered as a 'Big Data'.

■ Variety--How diverse are different types of data? It refers to nature of data that is structured, semi-structured and unstructured data. It also refers to heterogeneous sources.

■ Velocity--At what speed is new data generated? Velocity refers to the high speed of accumulation of data. In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc.

■ Veracity -- How accurate is the data? It refers to inconsistencies and uncertainty in data that is data which is available can sometimes get messy and quality and accuracy are difficult to control.

■ Value -- How effectively to transform a tsunami of data into business? Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information.

4. What are all the Challenges of Big Data?

The main challenges of Big Data are

i. Data capture - Data capture, or electronic data capture, is the process of extracting information from a document and converting it into data readable by a computer. ii. Curation - Data curation includes "all the processes needed for principled and controlled data creation, maintenance, and management, together with the capacity to add value to data".

iii. Storage - Data storage refers to magnetic, optical or mechanical media that records and preserves digital information for ongoing or future operations.

iv. Search - Searching is designed to check for an element/item or retrieve an element from any data storage.

v. Sharing - Data sharing is the practice of making data used for scholarly research available to other investigators

vi. Transfer - Data transfer refers to the secure exchange of large files between systems or organizations.

vii. Visualization - Data visualization is the graphical representation of information and data.

5. What are all the Benefits and uses/advantage of Data Science and Big Data Analytics?

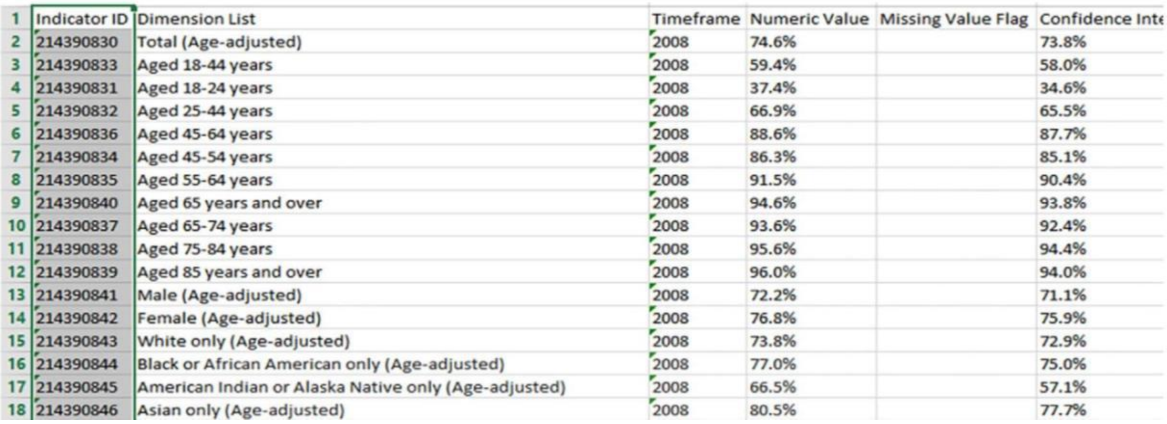
There are number of benefits/advantages while using the Big Data Analytics. Some of them are listed below.

- Commercial Companies in all business wish to analyses and gain insights into their customers, processes, staff, completion, and products. Many companies use data science to offer customers a better user experience, as well as to cross-sell, up-sell, and personalize their offerings.
- Human resource professionals use people analytics and text mining to screen candidates, monitor the mood of employees, and study informal networks among coworkers.
- Financial institutions use data science to predict stock markets, determine the risk of lending money, and learn how to attract new clients for their services.
- Many governmental organizations not only rely on internal data scientists to discover valuable information, but also share their data with the public. You can use this data to gain insights or build data-driven applications.
- Nongovernmental organizations (NGOs) can use it as a source for get funding. Many data scientists devote part of their time to helping NGOs, because NGOs often lack the resources to collect data and employ data scientists.
- Universities use data science in their research but also to enhance the study experience of their students. The rise of massive open online courses (MOOC) produces a lot of data, which allows universities to study how this type of learning can complement traditional classes.
- Data accumulation from multiple sources, including the Internet, social media platforms, online shopping sites, company databases, external third-party sources, etc.
- Real-time forecasting and monitoring of business as well as the market.
- Identify crucial points hidden within large datasets to influence business decisions.
- Promptly mitigate risks by optimizing complex decisions for unforeseen events and potential threats.

6. List the different types of Data used in Big Data Analytics and Data Science.

The major category of data types used in Big Data Analytics and Data Science are as follows.

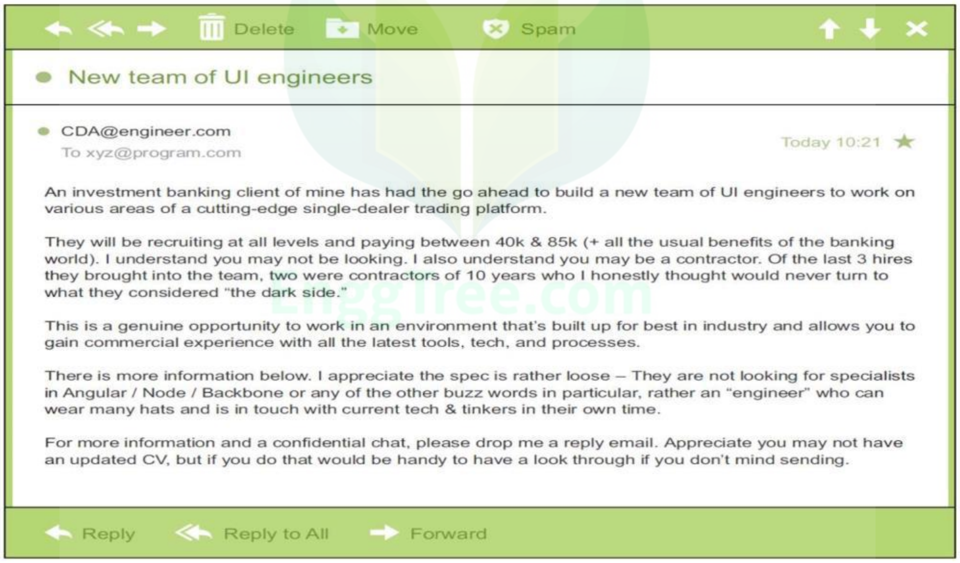
■ Structured - Structured data is data that depends on a data model and resides in a fixed field within a record. As such, it’s often easy to store structured data in tables within databases or Excel files. SQL, or Structured Query Language, is the preferred way to manage and query data that resides in databases.



1	Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Inte
2	214390830	Total (Age-adjusted)	2008	74.6%		73.8%
3	214390833	Aged 18-44 years	2008	59.4%		58.0%
4	214390831	Aged 18-24 years	2008	37.4%		34.6%
5	214390832	Aged 25-44 years	2008	66.9%		65.5%
6	214390836	Aged 45-64 years	2008	88.6%		87.7%
7	214390834	Aged 45-54 years	2008	86.3%		85.1%
8	214390835	Aged 55-64 years	2008	91.5%		90.4%
9	214390840	Aged 65 years and over	2008	94.6%		93.8%
10	214390837	Aged 65-74 years	2008	93.6%		92.4%
11	214390838	Aged 75-84 years	2008	95.6%		94.4%
12	214390839	Aged 85 years and over	2008	96.0%		94.0%
13	214390841	Male (Age-adjusted)	2008	72.2%		71.1%
14	214390842	Female (Age-adjusted)	2008	76.8%		75.9%
15	214390843	White only (Age-adjusted)	2008	73.8%		72.9%
16	214390844	Black or African American only (Age-adjusted)	2008	77.0%		75.0%
17	214390845	American Indian or Alaska Native only (Age-adjusted)	2008	66.5%		57.1%
18	214390846	Asian only (Age-adjusted)	2008	80.5%		77.7%

An Excel Table is an example for structured data

■ Unstructured: Unstructured data is data that isn't easy to fit into a data model because the content is context-specific or varying. One example of unstructured data is your regular email.



New team of UI engineers

CDA@engineer.com
To xyz@program.com

Today 10:21 ★

An investment banking client of mine has had the go ahead to build a new team of UI engineers to work on various areas of a cutting-edge single-dealer trading platform.

They will be recruiting at all levels and paying between 40k & 85k (+ all the usual benefits of the banking world). I understand you may not be looking. I also understand you may be a contractor. Of the last 3 hires they brought into the team, two were contractors of 10 years who I honestly thought would never turn to what they considered "the dark side."

This is a genuine opportunity to work in an environment that's built up for best in industry and allows you to gain commercial experience with all the latest tools, tech, and processes.

There is more information below. I appreciate the spec is rather loose – They are not looking for specialists in Angular / Node / Backbone or any of the other buzz words in particular, rather an "engineer" who can wear many hats and is in touch with current tech & tinkers in their own time.

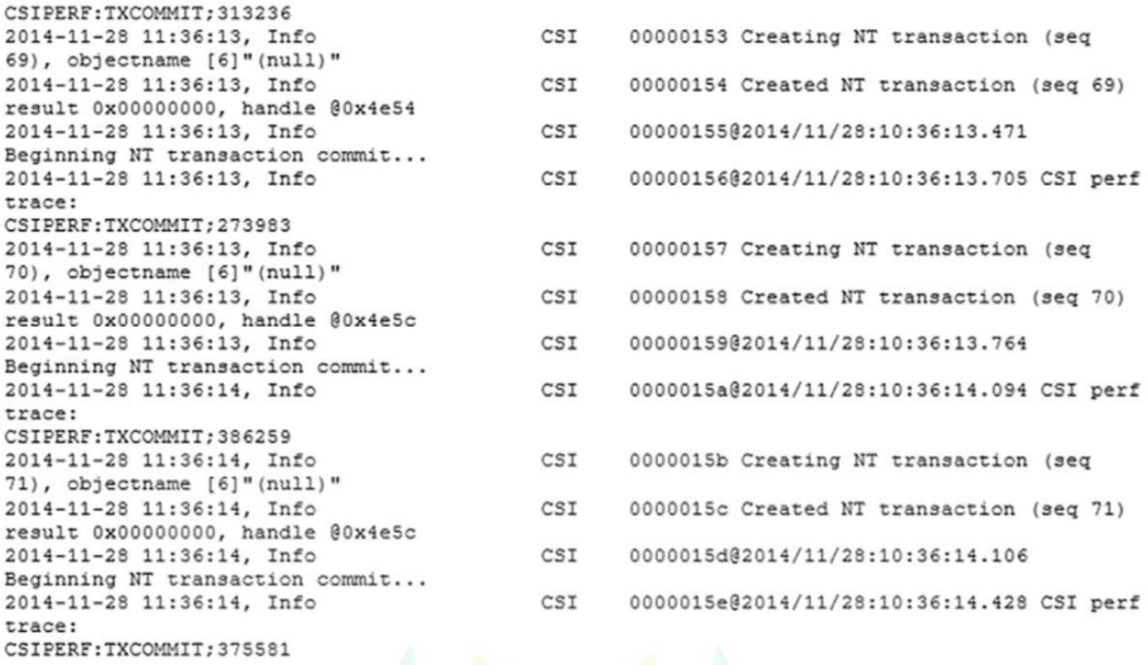
For more information and a confidential chat, please drop me a reply email. Appreciate you may not have an updated CV, but if you do that would be handy to have a look through if you don't mind sending.

Email is the best example for unstructured data and natural language data.

■ Natural language: Natural language is a special type of unstructured data; it's challenging to process because it requires knowledge of specific data science techniques and linguistics.

The natural language processing community has had success in entity recognition, topic recognition, summarization, text completion, and sentiment analysis, but models trained in one domain don't generalize well to other domains

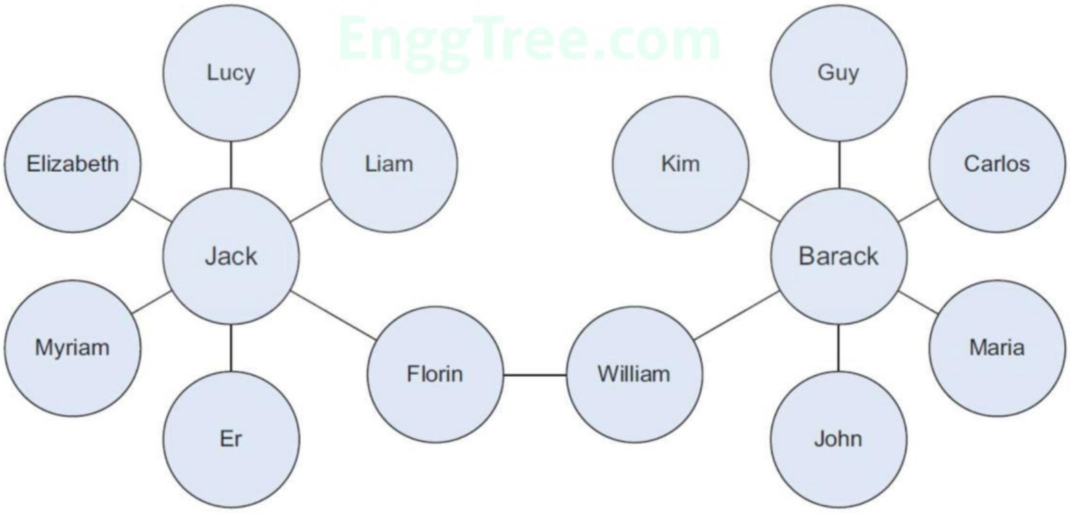
■ Machine-generated: Machine-generated data is information that's automatically created by a computer, process, application, or other machine without human intervention.



```
CSIPERF:TXCOMMIT;313236
2014-11-28 11:36:13, Info                CSI    00000153 Creating NT transaction (seq
69), objectname [6]"(null)"
2014-11-28 11:36:13, Info                CSI    00000154 Created NT transaction (seq 69)
result 0x00000000, handle @0x4e54
2014-11-28 11:36:13, Info                CSI    00000155@2014/11/28:10:36:13.471
Beginning NT transaction commit...
2014-11-28 11:36:13, Info                CSI    00000156@2014/11/28:10:36:13.705 CSI perf
trace:
CSIPERF:TXCOMMIT;273983
2014-11-28 11:36:13, Info                CSI    00000157 Creating NT transaction (seq
70), objectname [6]"(null)"
2014-11-28 11:36:13, Info                CSI    00000158 Created NT transaction (seq 70)
result 0x00000000, handle @0x4e5c
2014-11-28 11:36:13, Info                CSI    00000159@2014/11/28:10:36:13.764
Beginning NT transaction commit...
2014-11-28 11:36:14, Info                CSI    0000015a@2014/11/28:10:36:14.094 CSI perf
trace:
CSIPERF:TXCOMMIT;386259
2014-11-28 11:36:14, Info                CSI    0000015b Creating NT transaction (seq
71), objectname [6]"(null)"
2014-11-28 11:36:14, Info                CSI    0000015c Created NT transaction (seq 71)
result 0x00000000, handle @0x4e5c
2014-11-28 11:36:14, Info                CSI    0000015d@2014/11/28:10:36:14.106
Beginning NT transaction commit...
2014-11-28 11:36:14, Info                CSI    0000015e@2014/11/28:10:36:14.428 CSI perf
trace:
CSIPERF:TXCOMMIT;375581
```

Machine Generated Data

■ Graph-based or Network data: The graph structures use nodes, edges, and properties to represent and store graphical data. Graph-based data is a natural way to represent social networks, and its structure allows you to calculate specific metrics such as the influence of a person and the shortest path between two people



■ Audio, video, and images: Audio, image, and video are data types that pose specific challenges to a data scientist. Recently a company called DeepMind succeeded at creating an algorithm that's capable of learning how to play video games.

■ Streaming: While streaming data can take almost any of the previous forms, it has an extra property. The data flows into the system when an event happens instead of being loaded into a data store in a batch.

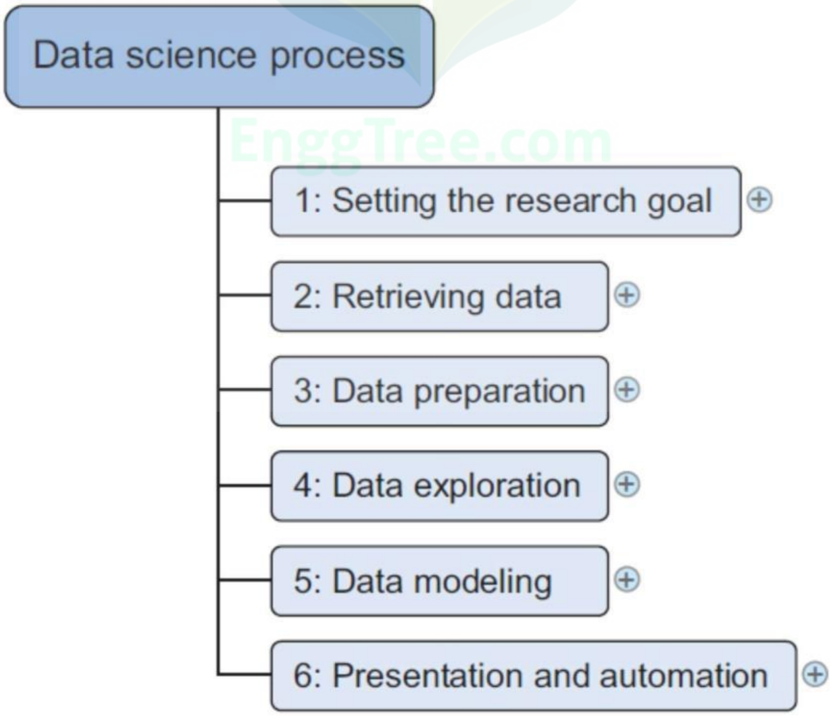
Examples are the “What’s trending” on Twitter, live sporting or music events, and the stock market.

7. Explain the data science process steps.

The data science process typically consists of six steps, it is shown in the following figure.

Setting the research goal: Data science is mostly applied in the context of an organization. To find the research goal and project character. The most important work in an organization is to do find out data value, and to identify, how the company benefits from that, what data and resources you need, a timetable, and deliverables.

Retrieving data: The second step is to collect data. In this step you ensure that you can use the data in your program, which means checking the existence of, quality, and access to the data. Data can also be delivered by third-party companies and takes many forms ranging from Excel spreadsheets to different types of databases.



Data Science Process

Data preparation: in this phase you enhance the quality of the data and prepare it for use in subsequent steps. This phase consists of three sub-phases: data cleansing removes false values from a data source and inconsistencies across data sources, data integration enriches data sources by combining information from multiple data sources, and data transformation ensures that the data is in a suitable format for use in your models.

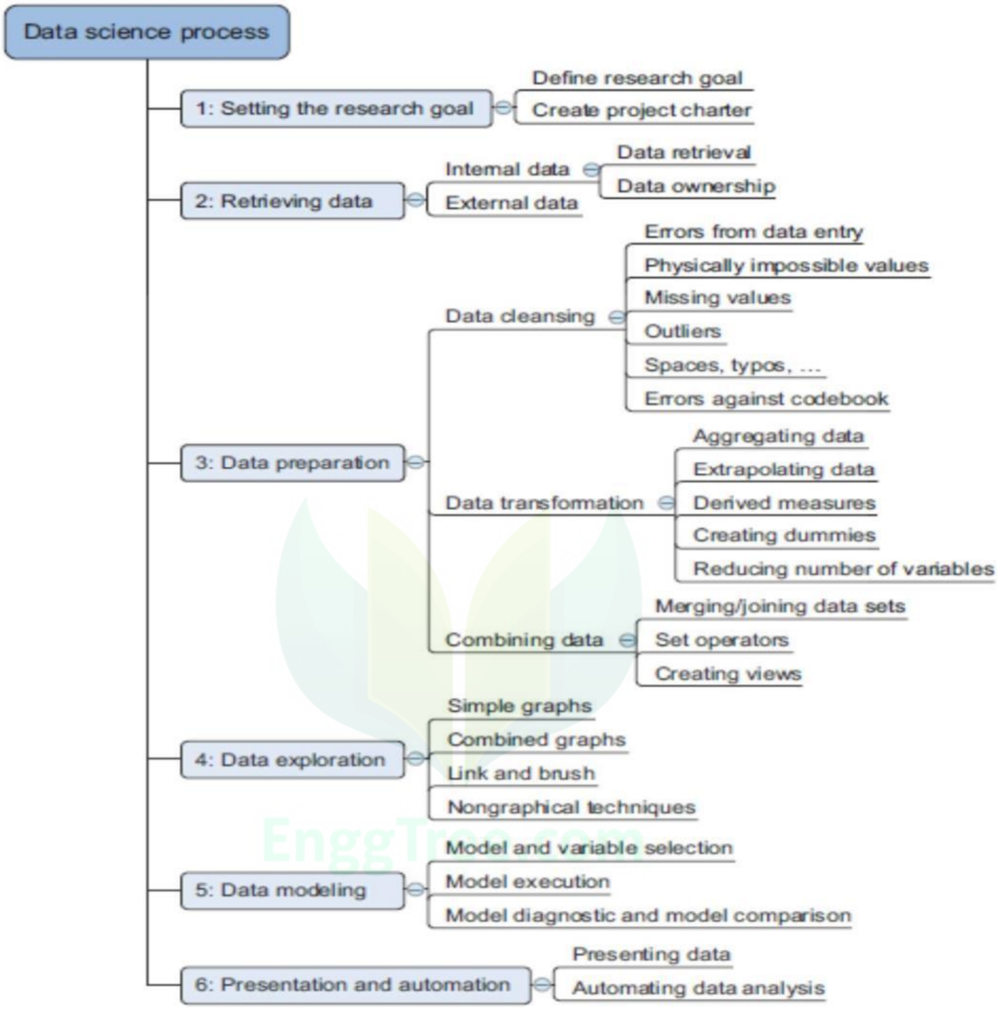
Data exploration: Data exploration is concerned with building a deeper understanding of your data. You try to understand how variables interact with each other, the distribution of the data, and whether there are outliers. To achieve this you mainly use descriptive statistics, visual techniques, and simple modeling. This step often goes by the abbreviation EDA, for Exploratory Data Analysis.

Data modeling or model building: In this phase you use models, domain knowledge, and insights about the data you found in the previous steps to answer the research question. You select a technique from the fields of statistics, machine learning, operations research, and so on. Building a model is an iterative process that involves selecting the variables for the model, executing the model, and model diagnostics.

Presentation and automation: Finally, you present the results to your business. These results can take many forms, ranging from presentations to research reports. Sometimes you’ll need to automate the execution of the process because the business will want to use the insights you gained in another project or enable an operational process to use the outcome from your model.

8. Explain the Data Science Process in detail.

Overview of the data science process



1. The first step of this process is setting a research goal. The main purpose here is making sure all the stakeholders understand the what, how, and why of the project. In every serious project this will result in a project charter.

2. The second phase is data retrieval. You want to have data available for analysis, so this step includes finding suitable data and getting access to the data from the data owner. The result is data in its raw form, which probably needs polishing and transformation before it becomes usable.

3. This step includes transforming the data from a raw form into data that's directly usable in your models. To achieve this, you'll detect and correct different kinds of errors in the data,

combine data from different data sources, and transform it. If you have successfully completed this step, you can progress to data visualization and modeling.

4. The fourth step is data exploration. The goal of this step is to gain a deep understanding of the data. You’ll look for patterns, correlations, and deviations based on visual and descriptive techniques. The insights you gain from this phase will enable you to start modeling.

5. Finally, we get to the sexiest part: model building (often referred to as “data modeling” throughout this book). It is now that you attempt to gain the insights or make the predictions stated in your project charter. If you’ve done this phase right, you’re almost done.

6. The last step of the data science model is presenting your results and automating the analysis, if needed. One goal of a project is to change a process and/or make better decisions. You may still need to convince the business that your findings will indeed change the business process as expected. This is where you can shine in your influencer role. The importance of this step is more apparent in projects on a strategic and tactical level. Certain projects require you to perform the business process over and over again, so automating the project will save time.

Step 1: Defining research goals and creating a project charter

- A project starts by understanding the what, the why, and the how of your project.
- What does the company expect you to do?
- Why does management place such a value on your research?
- The outcome should be a good understanding of the context, well-defined deliverables, and a plan of action with a timetable.
- Spend time understanding the goals and context of your research.
- Continue asking questions and devising examples until you grasp the exact business expectations, identify how your project fits in the bigger picture, appreciate how your research is going to change the business, and understand how they’ll use your results. □ Create a project charter

  After understanding the problems and goals, try to get a formal agreement on the deliverables.

  A project charter requires teamwork, and your input covers at least the following:

  ■ A clear research goal

  ■ The project mission and context

  ■ How you’re going to perform your analysis

  ■ What resources you expect to use

  ■ Proof that it’s an achievable project, or proof of concepts

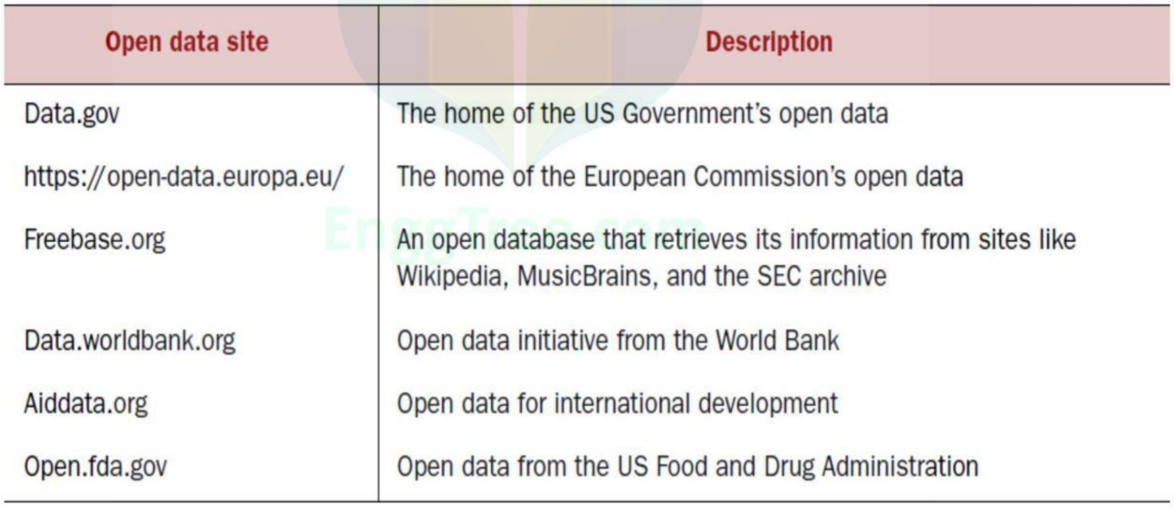
  ■ Deliverables and a measure of success

■ A timeline

Step 2: Retrieving data

- Data collection is the most important step.
- There are readymade data sets are available from Companies and organizations.
- Start with data stored within the company. Normally, it is stored in the Databases, Data marts, Data Warehouses, and Data Lakes.
- Knowledge of the data may be dispersed as people change positions and leave the company.
- Organizations understand the value and sensitivity of data.
- Often have policies in place so everyone has access to what they need.
- Don’t be afraid to shop around. There are number of companies are selling the data around the world.
- Many companies specialize in collecting valuable information. Nielsen and GFK are well known for this in the retail industry. Twitter, LinkedIn, and Facebook.

Table 2.1 A list of open-data providers that should get you started



Open data site	Description
Data.gov	The home of the US Government’s open data
https://open-data.europa.eu/	The home of the European Commission’s open data
Freebase.org	An open database that retrieves its information from sites like Wikipedia, MusicBrains, and the SEC archive
Data.worldbank.org	Open data initiative from the World Bank
Aiddata.org	Open data for international development
Open.fda.gov	Open data from the US Food and Drug Administration

- Do data quality checks now to prevent problems later.
- The retrieval of data is the first time you’ll inspect the data in the data science process.
- With data preparation, you do a more elaborate check, otherwise, those people are all using the dataset will be affected. You can use the statistical method to ensure the quality of data.
- During the exploratory phase your focus shifts to what you can learn from the data

Step 3: Cleansing, integrating, and transforming data

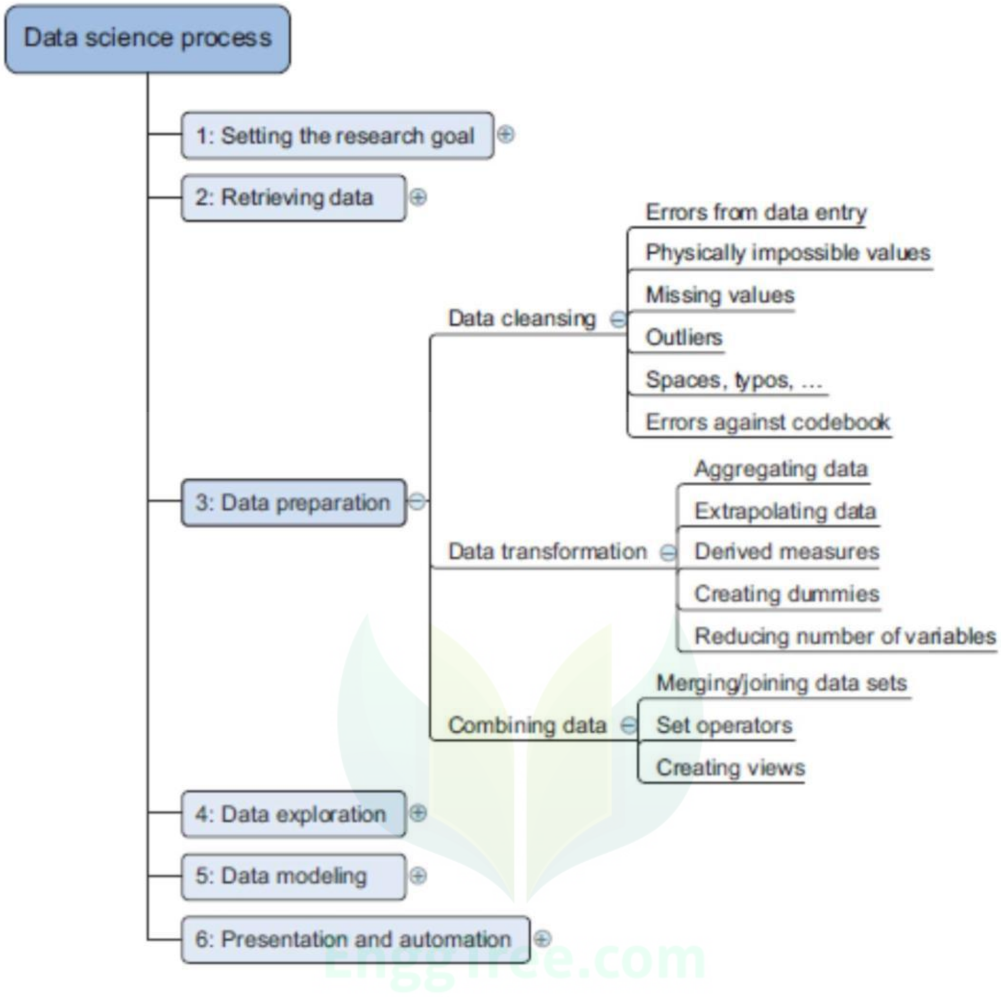
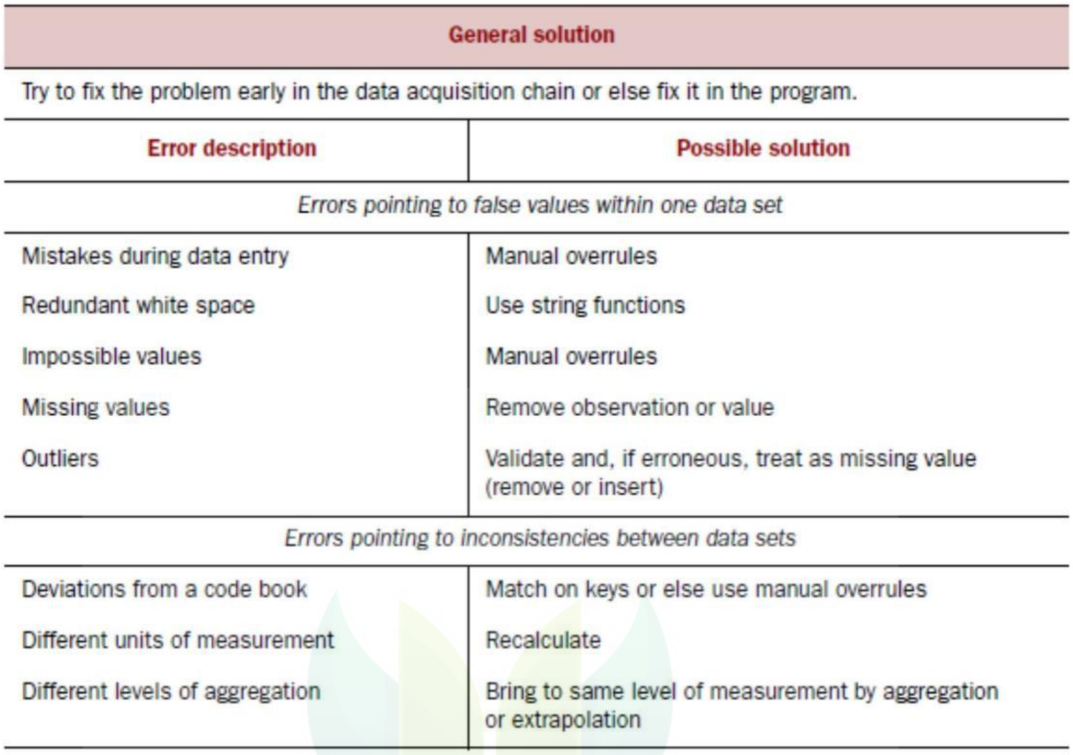


Figure 2.4 Step 3: Data preparation

 CLEANSING DATA

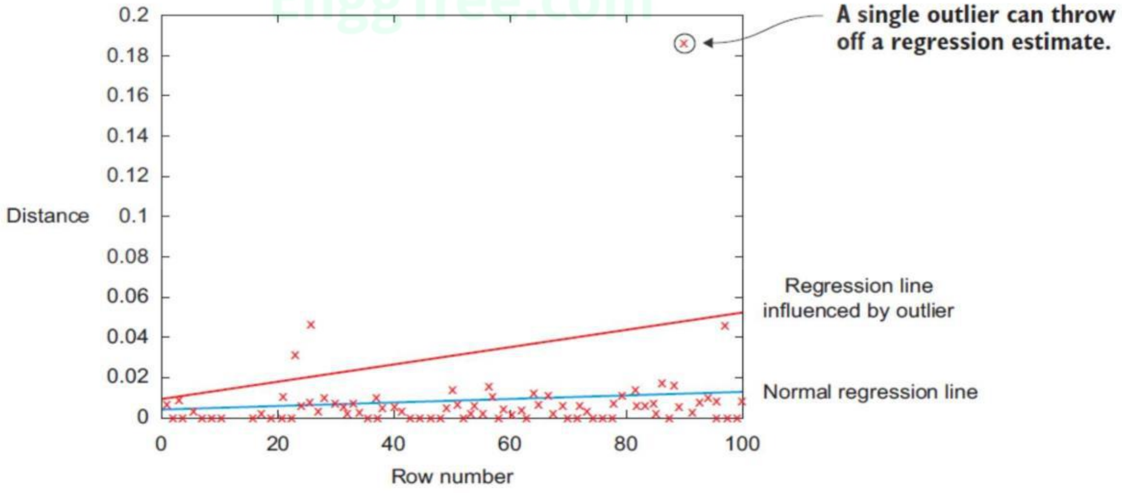
- Focuses on removing errors in your data so your data becomes a true and consistent representation of the processes it originates from.
- First type is the interpretation error. The common error types are given the following table.
- Second type of error points to inconsistencies between data sources.

Table 2.2 An overview of common errors



General solution	
Try to fix the problem early in the data acquisition chain or else fix it in the program.	
Error description	**Possible solution**
Errors pointing to false values within one data set	
Mistakes during data entry	Manual overrules
Redundant white space	Use string functions
Impossible values	Manual overrules
Missing values	Remove observation or value
Outliers	Validate and, if erroneous, treat as missing value (remove or insert)
Errors pointing to inconsistencies between data sets	
Deviations from a code book	Match on keys or else use manual overrules
Different units of measurement	Recalculate
Different levels of aggregation	Bring to same level of measurement by aggregation or extrapolation

- Sometimes a single observation has too much influence, this can point to an error in the data, but it can also be a valid point.



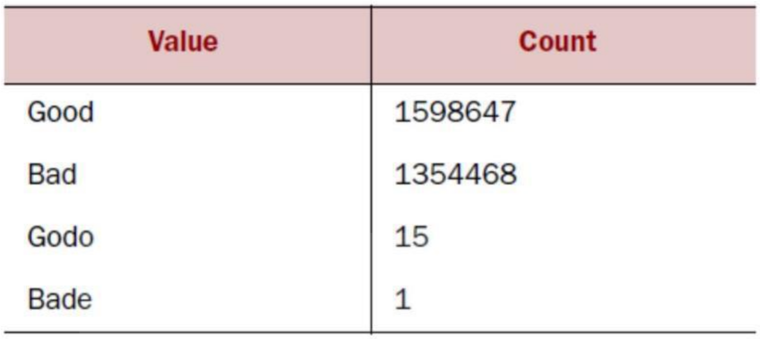
A single outlier can throw off a regression estimate.

Figure 2.5 The encircled point influences the model heavily and is worth investigating because it can point to a region where you don't have enough data or might indicate an error in the data, but it also can be a valid data point.

Data Entry Errors

Data collection and data entry are error-prone processes.

Table 2.3 Detecting outliers on simple variables with a frequency table



Value	Count
Good	1598647
Bad	1354468
Godo	15
Bade	1

Redundant Whitespace: At times we enter the redundant whitespace, this will lead complication in identification of strings. Most of the time string terminates with a whitespace.

Capital Letter Mismatches: Some time instead of using the capital letter, we will use the small letter, this is another problem in the data processing.

Impossible Values and Sanity Checks: It will accept the set of values in acceptable limits, it is essential to check such values in the real world. Example the age of a person cannot exceed 120. Therefore we need check such values.

Check = 1 <= age <= 120

Outliers

An observation that seems to be distant from other observations or, more specifically, one observation that follows a different logic or generative process than the other observations. The easiest way to find outliers is to use a plot or a table with the minimum and maximum values.

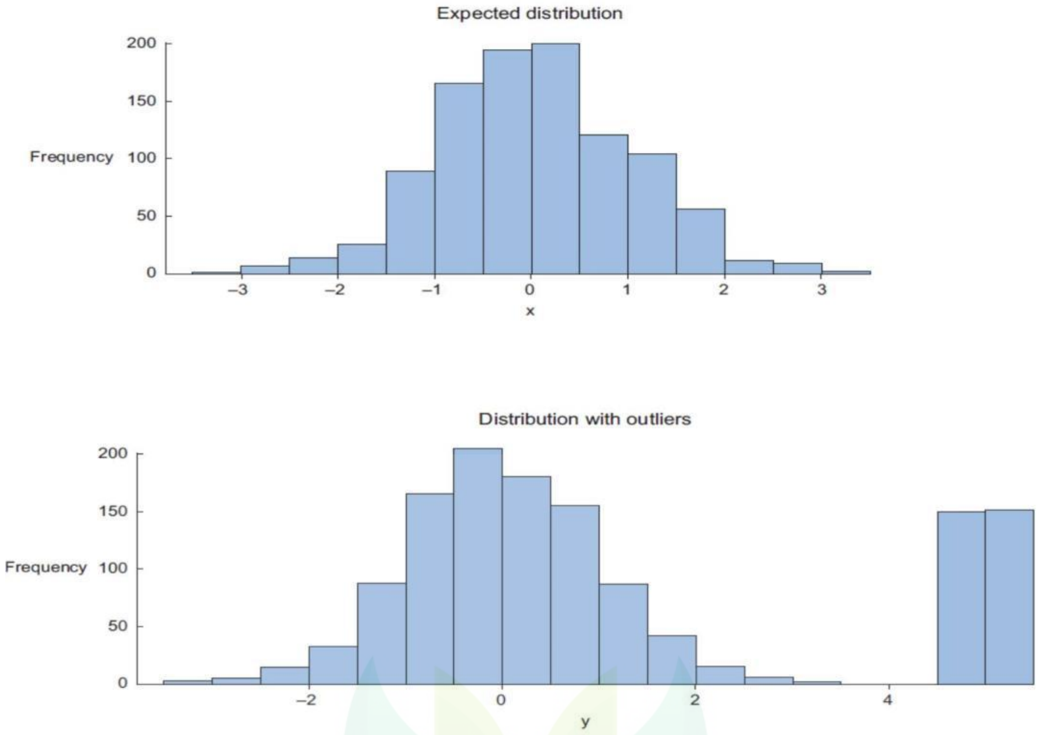
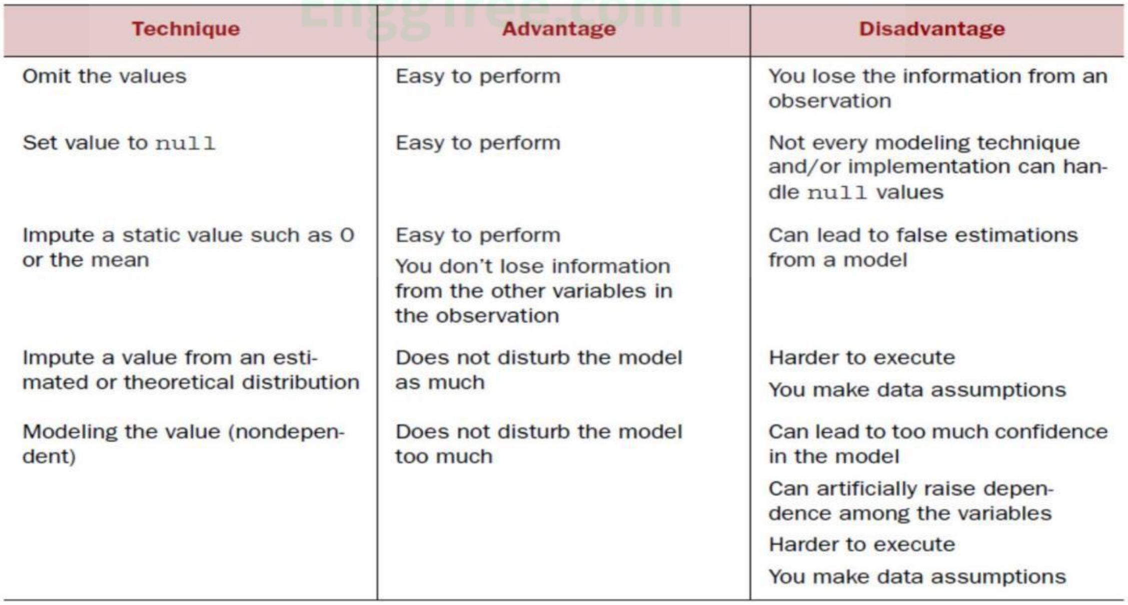


Figure 2.6 Distribution plots are helpful in detecting outliers and helping you understand the variable.

Dealing with Missing Values: Missing values need not be wrong, but still it need to be treated. The methods of treatment and their advantage and disadvantages are given in the following table.

Table 2.4 An overview of techniques to handle missing data



Technique	Advantage	Disadvantage
Omit the values	Easy to perform	You lose the information from an observation
Set value to `null`	Easy to perform	Not every modeling technique and/or implementation can handle `null` values
Impute a static value such as 0 or the mean	Easy to perform You don't lose information from the other variables in the observation	Can lead to false estimations from a model
Impute a value from an estimated or theoretical distribution	Does not disturb the model as much	Harder to execute You make data assumptions
Modeling the value (nondependent)	Does not disturb the model too much	Can lead to too much confidence in the model Can artificially raise dependence among the variables Harder to execute You make data assumptions

Deviations from a code book: A code book is a description of your data, a form of metadata. It contains things such as the number of variables per observation, the number of observations, and what each encoding within a variable means. One can use such a code book to correct the data, if it is missing or erroneous one.

Different Units of Measurement: Data set may be combined from different sources. Each source will use their own measurement. Therefore we need to look in carefully and correct / convert the same in a standard measure.

Different Levels of Aggregation: An example of this would be a data set containing data per week versus one containing data per work week. This type of error is generally easy to detect, and summarizing (or the inverse, expanding) the data sets will fix it. After cleaning the data errors, you combine information from different data sources.

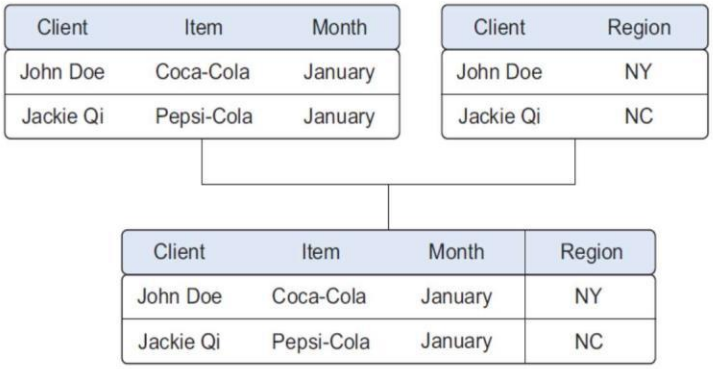
CORRECT ERRORS AS EARLY AS POSSIBLE

Data should be cleansed when acquired for many reasons:

- Decision Makers take important decisions.
- If errors are not corrected early on in the process, the cleansing will have to be done for every project that uses that data.
- Data errors may point to a business process that isn't working as designed.
- Data errors may point to defective equipment, such as broken transmission lines and defective sensors.
- Data errors can point to bugs in software or in the integration of software that may be critical to the company.

COMBINING DATA FROM DIFFERENT DATA SOURCES.

- Data is acquired from different sources and hence Data varies in size, type, and structure, ranging from databases and Excel files to text documents.
- The different ways of combining Data.
    - Joining Data from different tables.
    - Appending or stocking from different tables.
- An example for Joining Tables is shown below. The key uniquely identified is called as Primary key, which is used for joining data and elimination of redundancy of data.



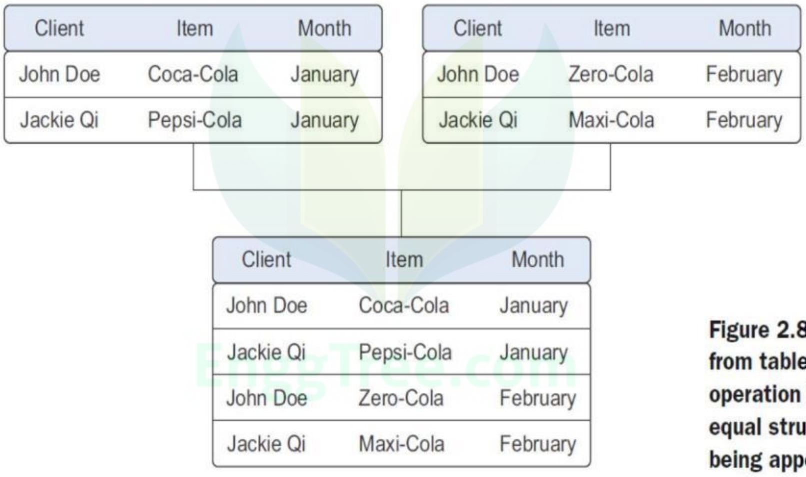
Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January

Client	Region
John Doe	NY
Jackie Qi	NC

Client	Item	Month	Region
John Doe	Coca-Cola	January	NY
Jackie Qi	Pepsi-Cola	January	NC

Figure 2.7 Joining two tables on the Item and Region keys

- An example for Appending or Stacking Data from different Tables. In general, the SQL query is used for appending the or stacking the tables.



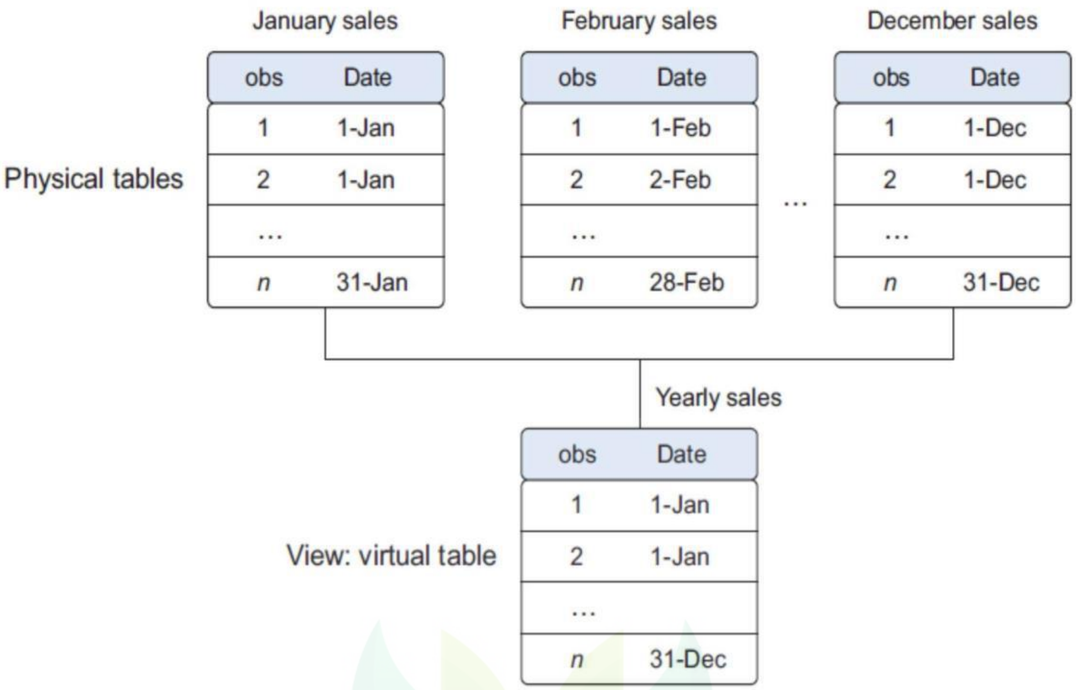
Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January

Client	Item	Month
John Doe	Zero-Cola	February
Jackie Qi	Maxi-Cola	February

Client	Item	Month
John Doe	Coca-Cola	January
Jackie Qi	Pepsi-Cola	January
John Doe	Zero-Cola	February
Jackie Qi	Maxi-Cola	February

Figure 2.8 Appending data from tables is a common operation but requires an equal structure in the tables being appended.

- Sometimes the Join or Appending is harder due to the disk space restrictions. In order to avoid this problem, the Views are used to Join or append the tables. It will only create the logical view, without physical creation of data. This is shown in the following table.



January sales

obs	Date
1	1-Jan
2	1-Jan
…	
n	31-Jan

February sales

obs	Date
1	1-Feb
2	2-Feb
…	
n	28-Feb

December sales

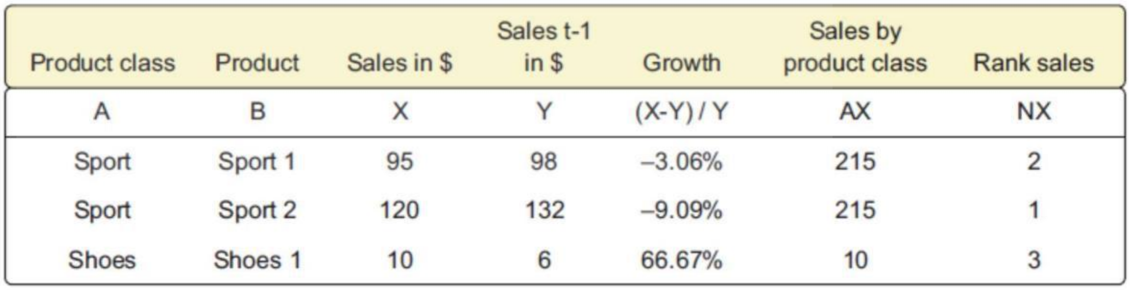
obs	Date
1	1-Dec
2	1-Dec
…	
n	31-Dec

Yearly sales

obs	Date
1	1-Jan
2	1-Jan
…	
n	31-Dec

Figure 2.9 A view helps you combine data without replication.

- o Enriching Aggregated Measures.
- o Extra measures such as these can add perspective. Looking at figure 2.10, we now have an aggregated data set, which in turn can be used to calculate the participation of each product within its category. This could be useful during data exploration but more so when creating data models.



Product class	Product	Sales in $	Sales t-1 in $	Growth	Sales by product class	Rank sales
A	B	X	Y	(X-Y) / Y	AX	NX
Sport	Sport 1	95	98	–3.06%	215	2
Sport	Sport 2	120	132	–9.09%	215	1
Shoes	Shoes 1	10	6	66.67%	10	3

Figure 2.10 Growth, sales by product class, and rank sales are examples of derived and aggregate measures.

TRANSFORMING DATA

- o Transforming data into suitable model
- o Transforming the data into suitable to the model is essential. This helps to identify the relationship among the data set.

x	1	2	3	4	5	6	7	8	9	10
log(x)	0.00	0.43	0.68	0.86	1.00	1.11	1.21	1.29	1.37	1.43
y	0.00	0.44	0.69	0.87	1.02	1.11	1.24	1.32	1.38	1.46

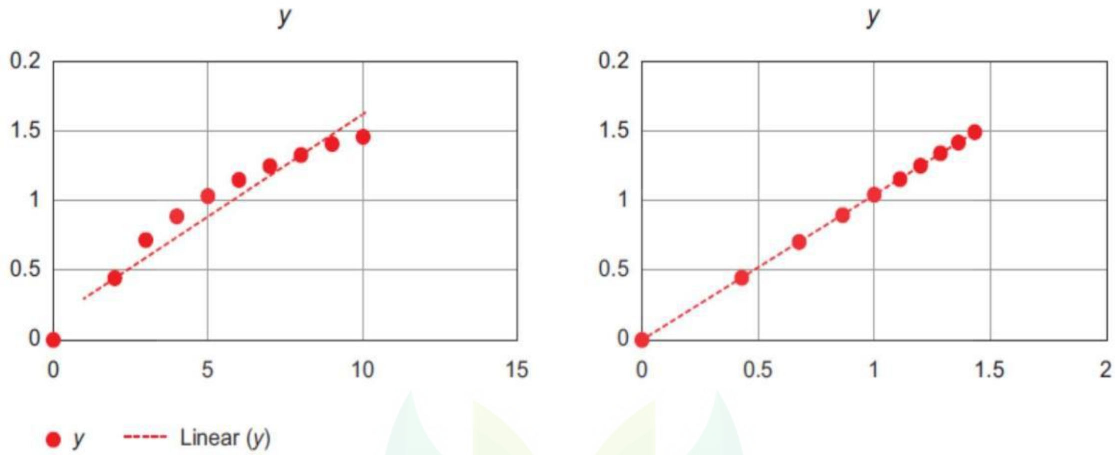


Figure 2.11 Transforming x to log x makes the relationship between x and y linear (right), compared with the non-log x (left).

o In the above example, we just want to transform the Y = ae^{bx} data into linear model by taking the log x value. This is essential in some cases. o Reducing the number of variable.

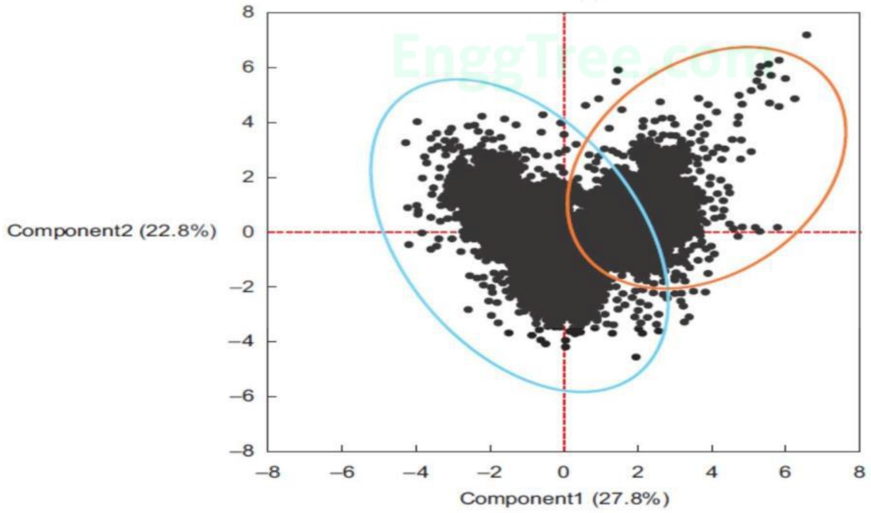
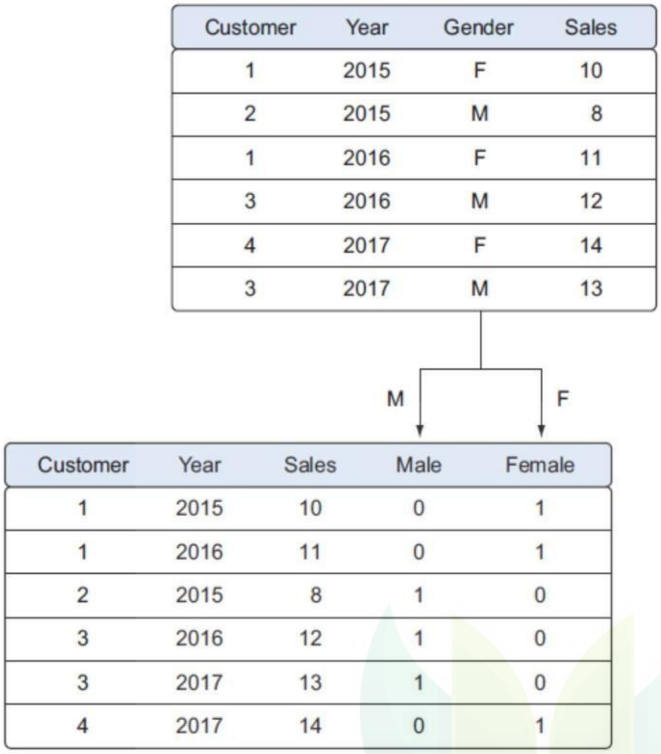


Figure 2.12 Variable reduction allows you to reduce the number of variables while maintaining as much information as possible.

- o In some cases it is essential to identify the most important attributes and select those value for analysis. The PCA (Principal Component Analysis) used for this purpose, which will avoid the unessential variables.
- o Turning Variables into Dummies.
  Sometimes it is essential to transform the dataset into binary values to avoid the processing difficulties and this mechanism is known as dummies.



Customer	Year	Gender	Sales
1	2015	F	10
2	2015	M	8
1	2016	F	11
3	2016	M	12
4	2017	F	14
3	2017	M	13

Customer	Year	Sales	Male	Female
1	2015	10	0	1
1	2016	11	0	1
2	2015	8	1	0
3	2016	12	1	0
3	2017	13	1	0
4	2017	14	0	1

Figure 2.13 Turning variables into dummies is a data transformation that breaks a variable that has multiple classes into multiple variables, each having only two possible values: 0 or 1.

Step 4: Exploratory Data Analysis

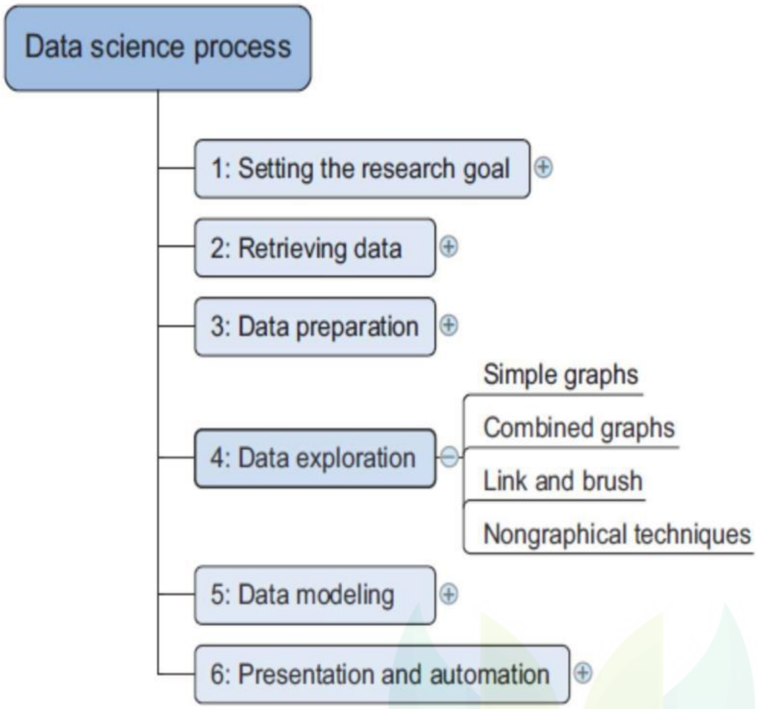


Figure 2.14 Step 4: Data exploration

The exploratory analysis are the methods used to understand the trend, relationship among the variables. This could be performed by means for different methods as given below. Information becomes much easier to grasp when shown in a picture, therefore you mainly use graphical techniques to gain an understanding of your data and the interactions between variables.

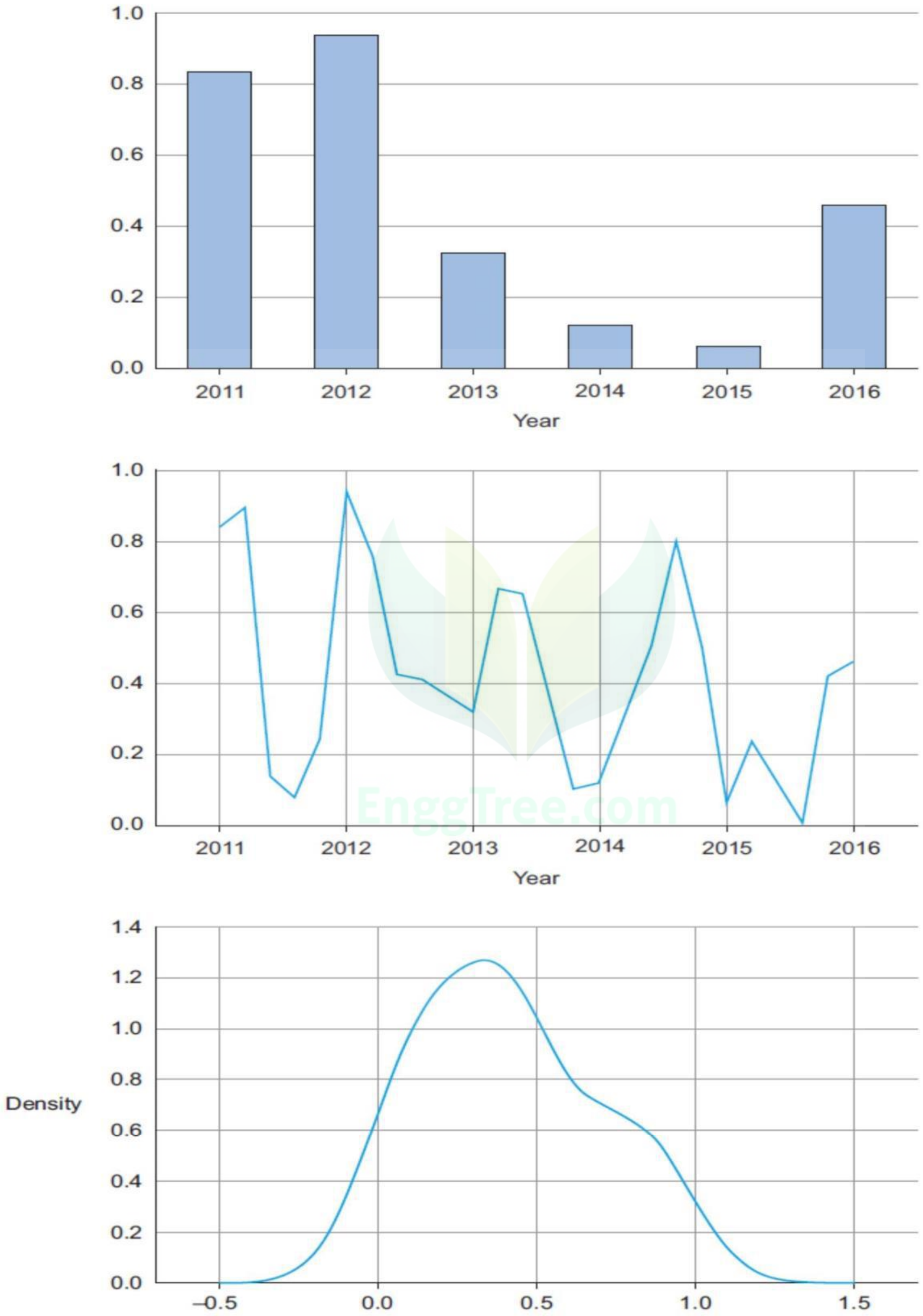


Figure 2.15 From top to bottom, a bar chart, a line plot, and a distribution are some of the graphs used in exploratory analysis.

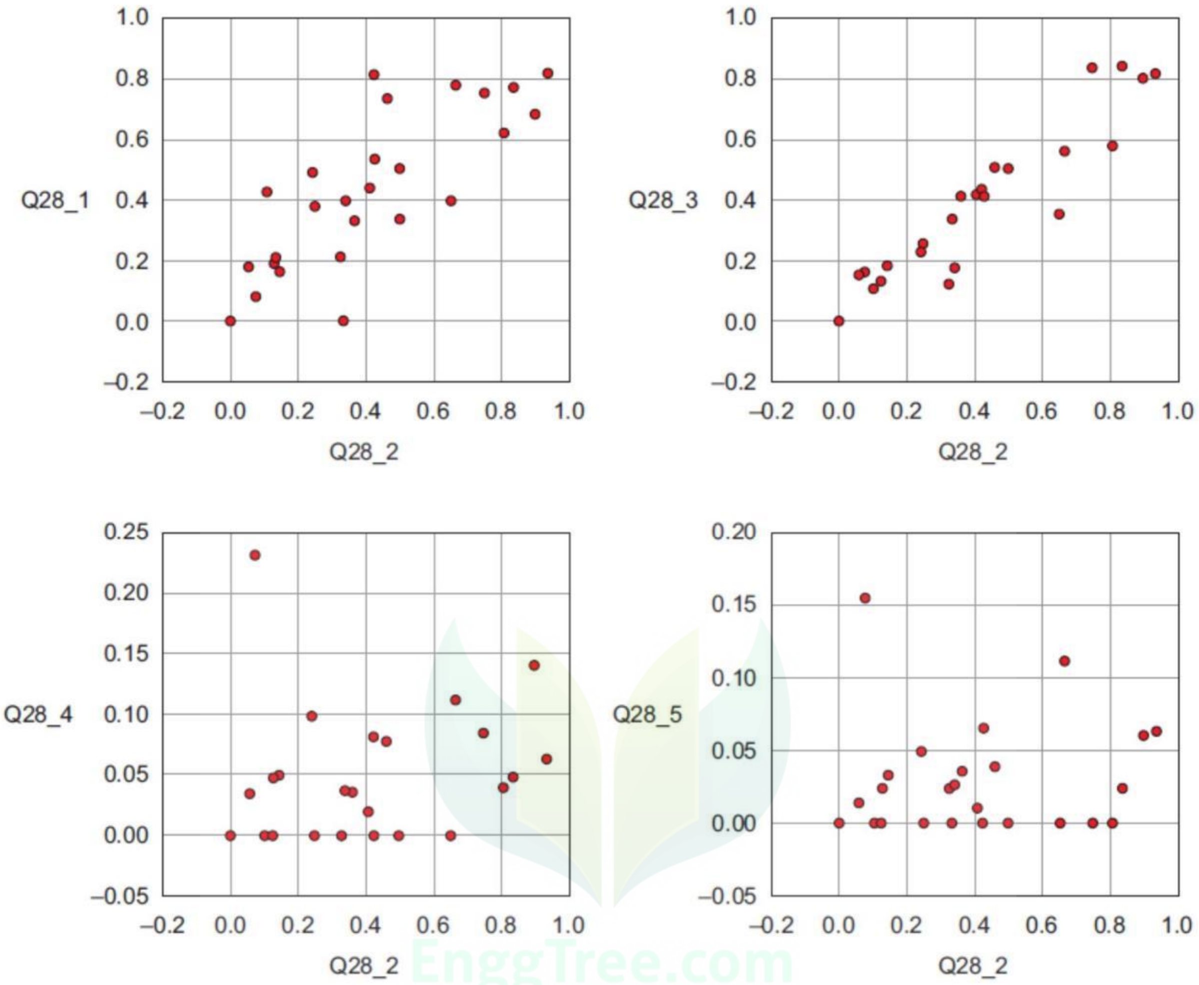


Figure 2.16 Drawing multiple plots together can help you understand the structure of your data over multiple variables.

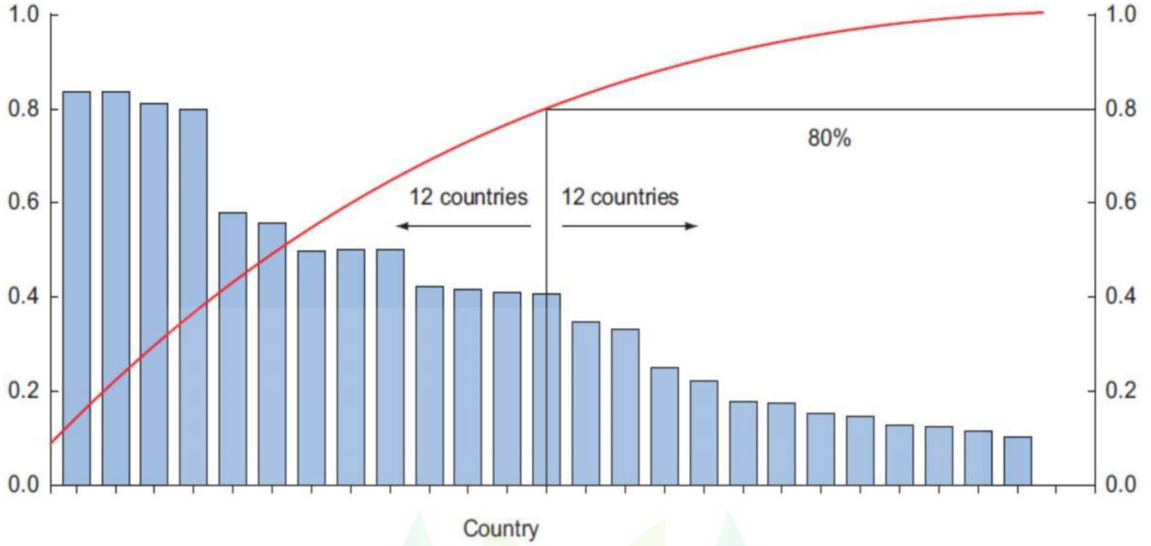


Figure 2.17 A Pareto diagram is a combination of the values and a cumulative distribution. It's easy to see from this diagram that the first 50% of the countries contain slightly less than 80% of the total amount. If this graph represented customer buying power and we sell expensive products, we probably don't need to spend our marketing budget in every country; we could start with the first 50%.

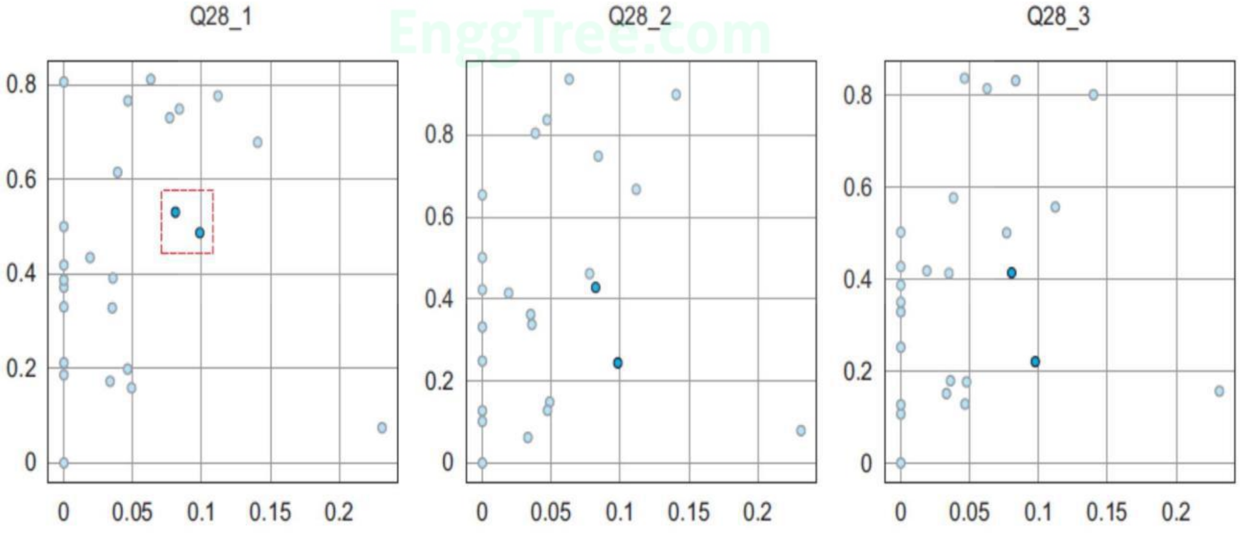


Figure 2.18 Link and brush allows you to select observations in one plot and highlight the same observations in the other plots.

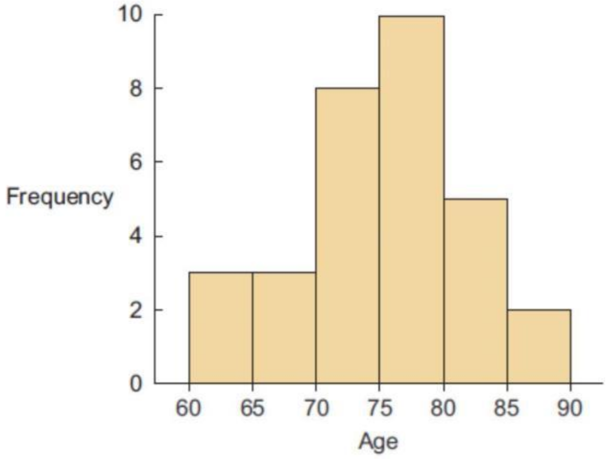


Figure 2.19 Example histogram: the number of people in the age-groups of 5-year intervals

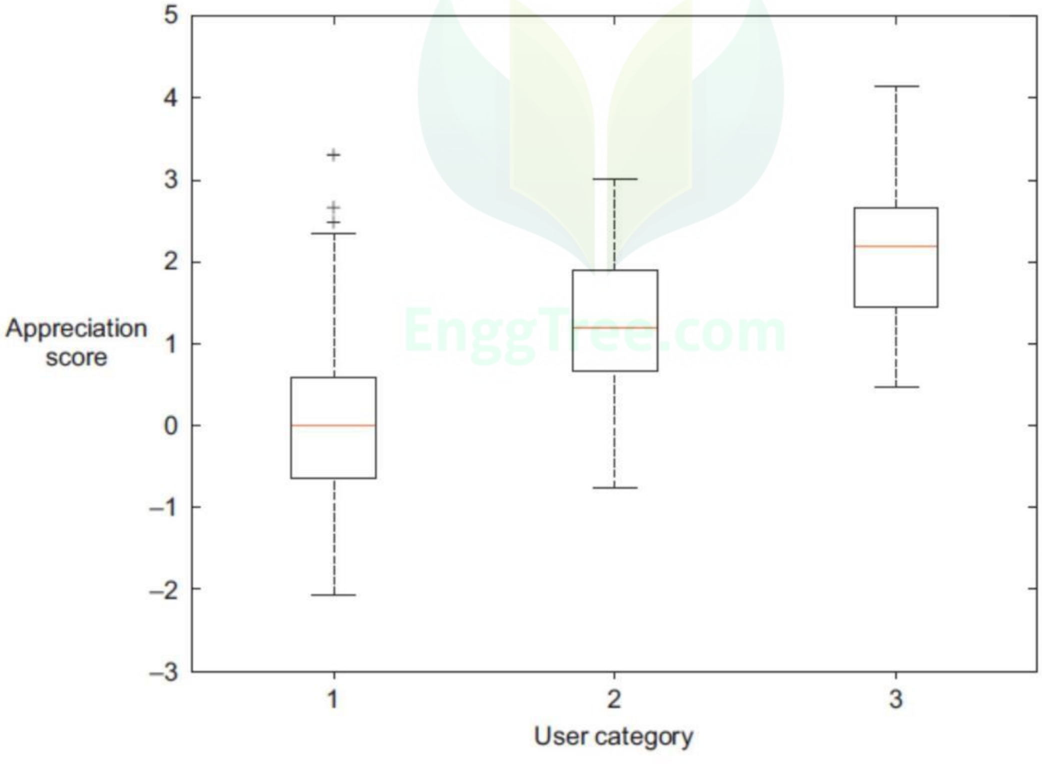


Figure 2.20 Example boxplot: each user category has a distribution of the appreciation each has for a certain picture on a photography website.

Step 5: Build the models

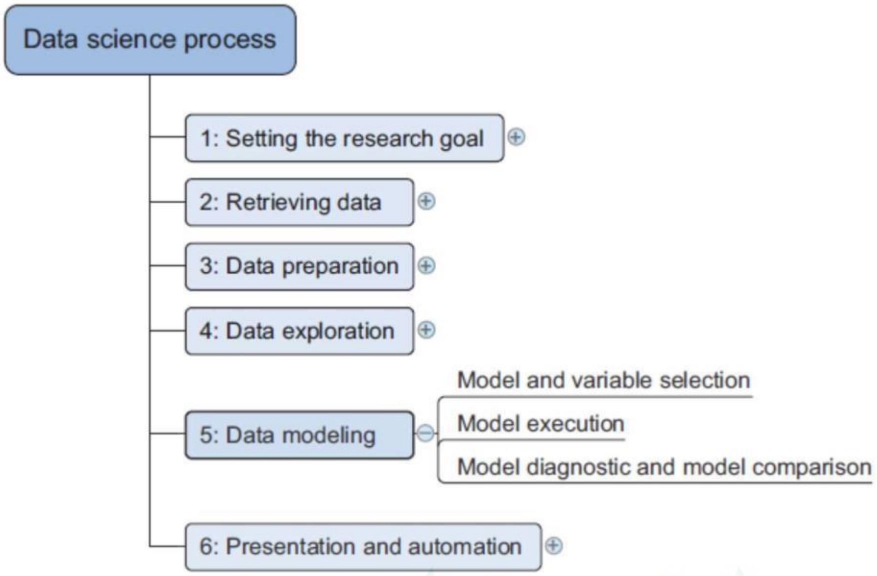

Data science process
1: Setting the research goal
2: Retrieving data
3: Data preparation
4: Data exploration
5: Data modeling
Model and variable selection
Model execution
Model diagnostic and model comparison
6: Presentation and automation

Figure 2.21 Step 5: Data modeling

It is a three step process

1. Selection of a modeling technique and variables to enter in the model
2. Execution of the model
3. Diagnosis and model comparison

Building a model is an iterative process that involves selecting the variables for the model, executing the model, and model diagnostics.

Model and variable selection

In order to select the variable we should ask the following questions yourself and do the variable selection.

- Must the model be moved to a production environment and, if so, would it be easy to implement?
- How difficult is the maintenance on the model: how long will it remain relevant if left untouched? o Does the model need to be easy to explain?

Model Execution

There are number mechanisms are available from both statistical and machine learning domain.

First one is the linear Regression: This is the normal line fitting mechanism to extrapolate the value or doing the prediction. Normally, we solve the equation y = mx+b and try to find the value of m and b with respect to the given dataset.

Listing 2.1 Executing a linear prediction model on semi-random data

```
import statsmodels.api as sm
import numpy as np
predictors = np.random.random(1000).reshape(500,2)
target = predictors.dot(np.array([0.4, 0.6])) + np.random.random(500)
lmRegModel = sm.OLS(target,predictors)
result = lmRegModel.fit()
result.summary()
```

Imports required Python modules.

Creates random data for predictors (x-values) and semi-random data for the target (y-values) of the model. We use predictors as input to create the target so we infer a correlation here.

Fits linear regression on data.

Shows model fit statistics.

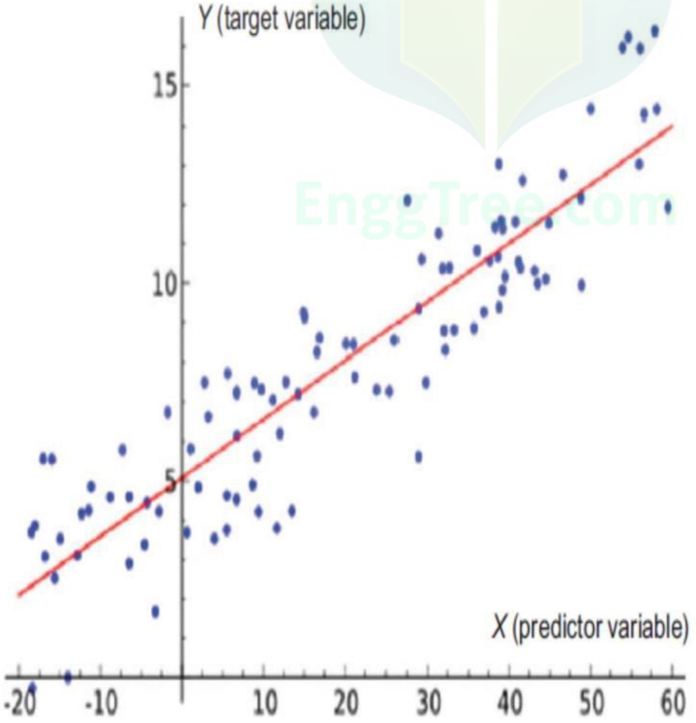
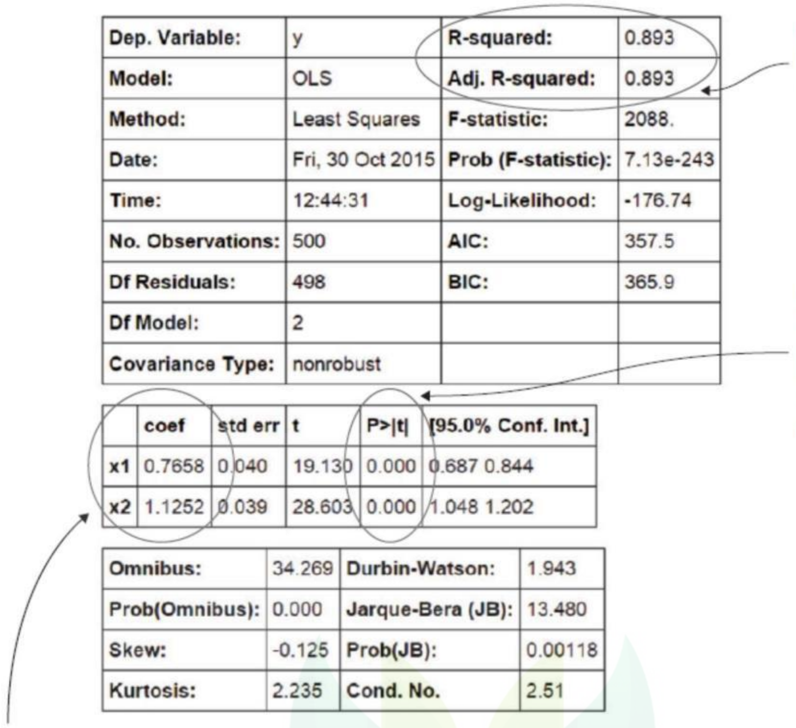


Figure 2.22 Linear regression tries to fit a line while minimizing the distance to each point



Dep. Variable:	y	R-squared:	0.893
Model:	OLS	Adj. R-squared:	0.893
Method:	Least Squares	F-statistic:	2088.
Date:	Fri, 30 Oct 2015	Prob (F-statistic):	7.13e-243
Time:	12:44:31	Log-Likelihood:	-176.74
No. Observations:	500	AIC:	357.5
Df Residuals:	498	BIC:	365.9
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
x1	0.7658	0.040	19.130	0.000	0.687 0.844
x2	1.1252	0.039	28.603	0.000	1.048 1.202

Omnibus:	34.269	Durbin-Watson:	1.943
Prob(Omnibus):	0.000	Jarque-Bera (JB):	13.480
Skew:	-0.125	Prob(JB):	0.00118
Kurtosis:	2.235	Cond. No.	2.51

Model fit: higher is better but too high is suspicious.

p-value to show whether a predictor variable has a significant influence on the target. Lower is better and <0.05 is often considered "significant."

Linear equation coefficients. y = 0.7658xl + l.l252x2.

Figure 2.23 Linear regression model information output

Model fit: It is essential to ensure the fitting of the model with respect to the given data. In order to do that one we find the difference between the R-Square value and Adj. R-Square value should be minimal. If it is minimal then it is considered to be the suitable model for given dataset.

Predictor Variable have coefficients. It will change because of adding more sample into the system.

Predictor significance in the target variable p. It should be less than 0.5. It is another factor which study the impact of the model in the prediction.

Second one for the Classification: K-Nearest Neighbor is the one of the supervised classification mechanism. In this classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small).

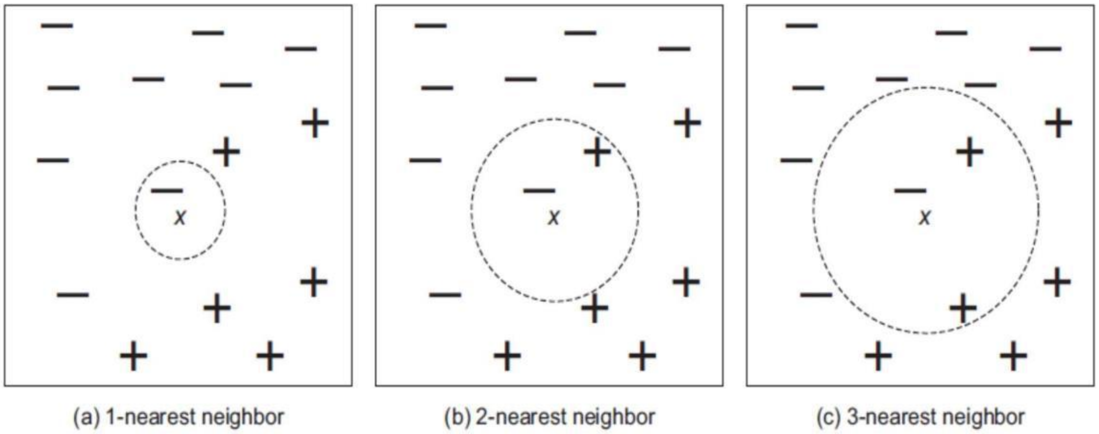
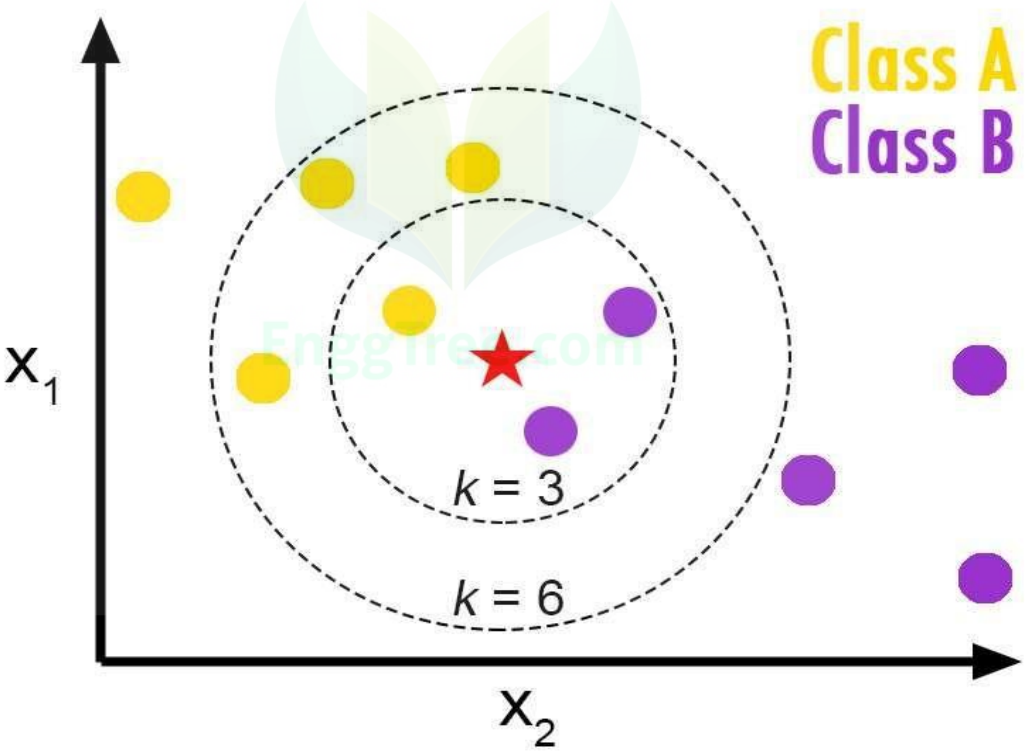
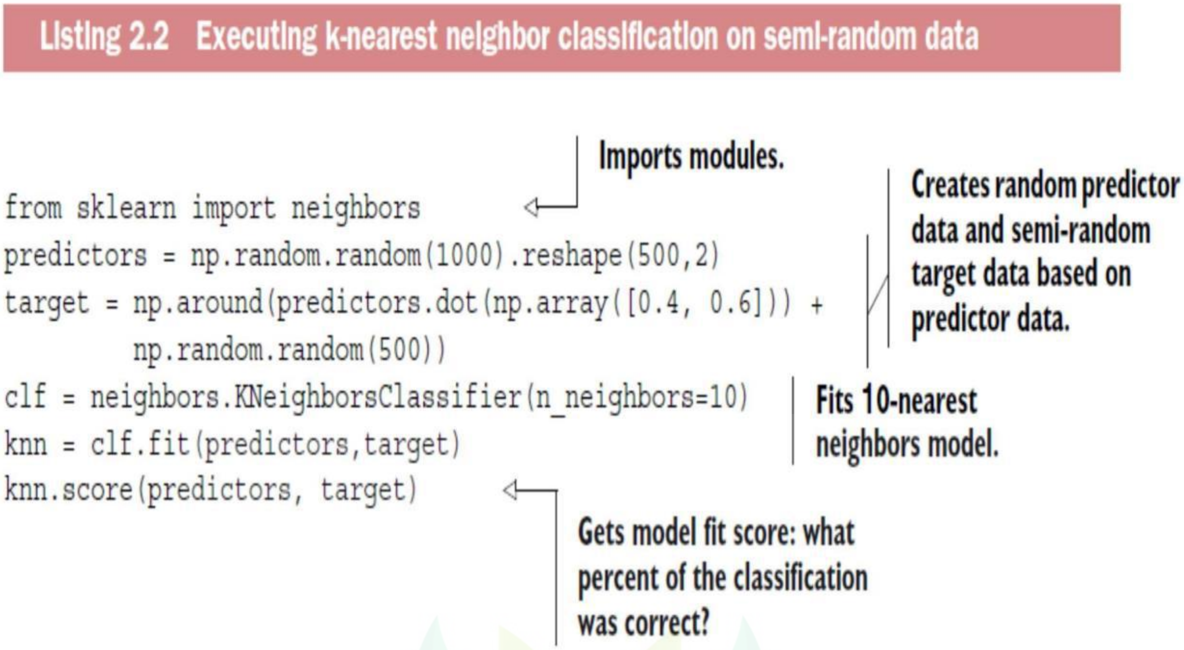


Figure 2.24 K-nearest neighbor techniques look at the k-nearest point to make a prediction.





Listing 2.2 Executing k-nearest neighbor classification on semi-random data

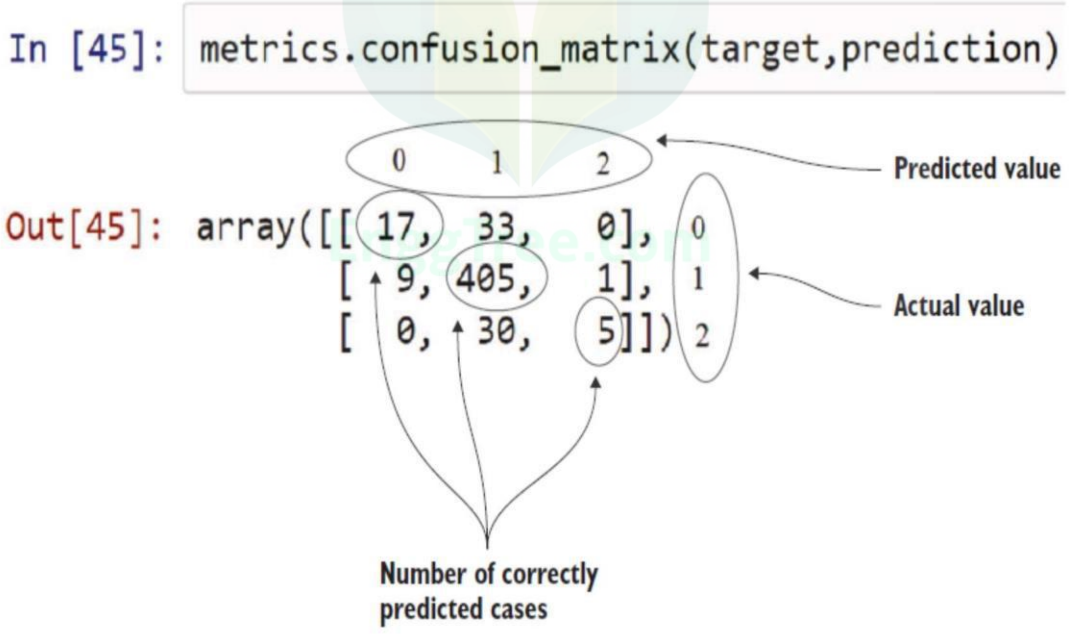
```
from sklearn import neighbors
predictors = np.random.random(1000).reshape(500,2)
target = np.around(predictors.dot(np.array([0.4, 0.6])) +
         np.random.random(500))
clf = neighbors.KNeighborsClassifier(n_neighbors=10)
knn = clf.fit(predictors,target)
knn.score(predictors, target)
```

Imports modules.

Creates random predictor data and semi-random target data based on predictor data.

Fits 10-nearest neighbors model.

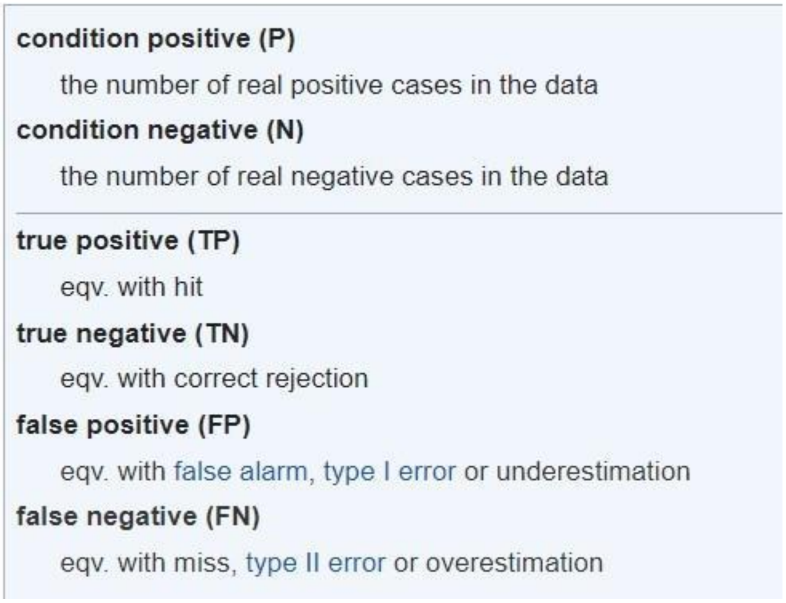
Gets model fit score: what percent of the classification was correct?



```
In [45]: metrics.confusion_matrix(target,prediction)

Out[45]: array([[ 17,  33,   0],
                [  9, 405,   1],
                [  0,  30,   5]])
```

Figure 2.25 Confusion matrix: it shows how many cases were correctly classified and incorrectly classified by comparing the prediction with the real values. Remark: the classes (0,1,2) were added in the figure for clarification.

o It is an error matrix to represent the accuracy of prediction.



condition positive (P)

the number of real positive cases in the data

condition negative (N)

the number of real negative cases in the data

true positive (TP)

eqv. with hit

true negative (TN)

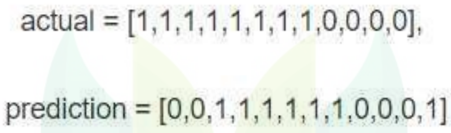
eqv. with correct rejection

false positive (FP)

eqv. with false alarm, type I error or underestimation

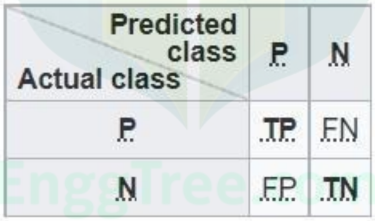
false negative (FN)

eqv. with miss, type II error or overestimation

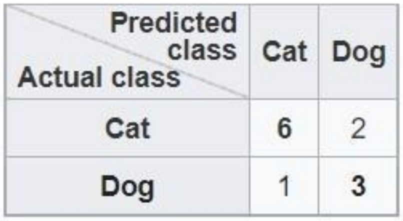


actual = [1,1,1,1,1,1,1,1,0,0,0,0],

prediction = [0,0,1,1,1,1,1,1,0,0,0,1]



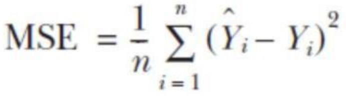
Predicted class / Actual class	P	N
P	TP	FN
N	FP	TN



Predicted class / Actual class	Cat	Dog
Cat	6	2
Dog	1	3

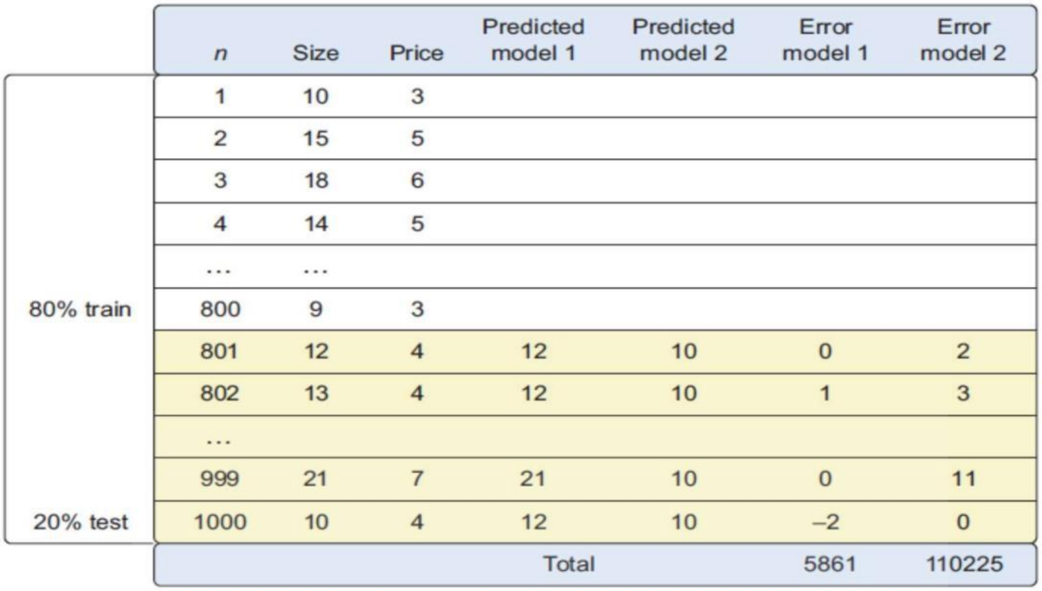
Model diagnostics and model comparison

- Mean Square Error is the standard measurement used to predict the accuracy of the system. If it is small then accuracy of the classification model is high.



$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

Figure 2.26 Formula for mean square error



	n	Size	Price	Predicted model 1	Predicted model 2	Error model 1	Error model 2
	1	10	3				
	2	15	5				
	3	18	6				
	4	14	5				
	…	…					
80% train	800	9	3				
	801	12	4	12	10	0	2
	802	13	4	12	10	1	3
	…						
	999	21	7	21	10	0	11
20% test	1000	10	4	12	10	–2	0
				Total		5861	110225

Figure 2.27 A holdout sample helps you compare models and ensures that you can generalize results to data that the model has not yet seen.

Step 6: Presenting findings and building applications on top of them.

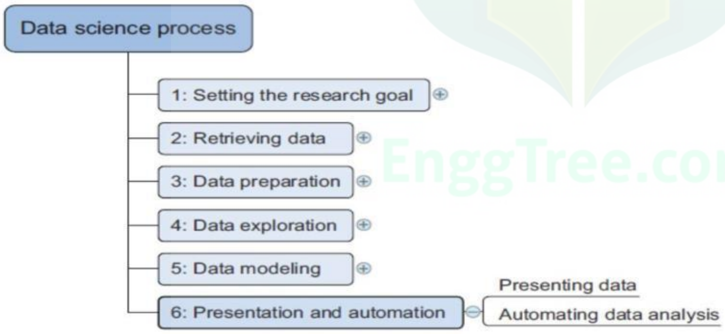


Figure 2.28 Step 6: Presentation and automation

- o Present the visualization as per your requirements. o The visualization techniques are discussed in the exploratory analysis in detail. o Change the visualization as per changes or modifications in the dataset automatically.

Inferential Statistics-II

Why Hypothesis Test?

The variability among sample means must be considered when we attempt to decide whether the observed difference b/w 533 and 500 is real or merely traneitory.

Importance of standard error:-

When expressed as z, the ratio of the observed diff., to the standard error is small enough to be reasonably attributed to chance, we retain Ho. otherwise, if the ratio of the observed difference to the standard error is too large to be reasonably attributed to chance, as in the SAT example, we reject Ho.

Befone generalizing beyond the existing data, we must measure the effect of chance; that we should obtain a value for the std., error.

for SAT example, Increase $\sigma_{\bar{X}}$ from 11 to 33 and note that even though the observed difference remains the same (533-500), we could retain Ho b'coz now z' would equal '1'.

$$z_0 = \frac{533-500}{33} = \frac{33}{33} = 1 \quad ; \quad z_c = 1.96$$

$z_0 < z_c \Rightarrow$ Condition satisfied so we retain Ho

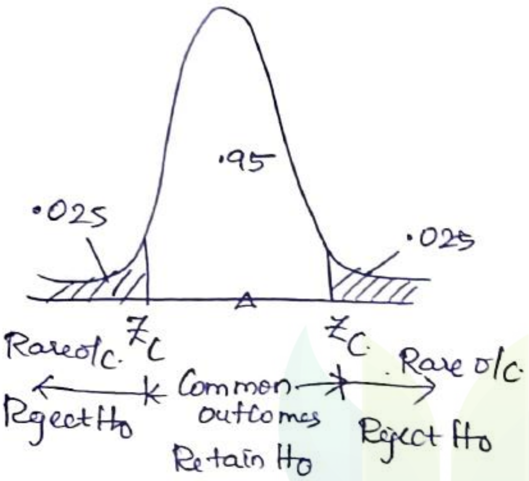
Possibility of Incorrect Decisions

Type-I error or false alarm.

- rejecting a true H_0.

Type-II error or Miss

- retaining a false H_0



Strong or Weak Decisions:

Retaining H_0 is a weak Decision

- It is difficult to retain H_0 & reject H_0.

→ H_0 is retained if 'Z_0' qualifies as common outcome. & so H_0 retained.

→ However, the same observed result also would qualify as a common outcome when the Orginal value H_0(500) is replaced with a slightly different value. If so retaining H_0 is weak decision.

→ Statisticians describe it as failure to reject H_0 rather than retention of H_0.

Two types of Errors in hypothesis Testing :

Four possibilities.

i) The hypothesis is true but ouf test rejects it. (Type 1 error)

ii) The hypothesis is false but our test accepts it. (Type 2 error)

iii) Hyp - True, acept test

iv) Hyp - false, reject test.

Type 1 error - repnesented by 'α'

$$\alpha = \text{Prob}(\text{Type 1 error})$$
$$= \text{Prob}(\text{Rejecting } H_0/H_a \text{ is True})$$

Type 2 error - represented by 'β'

$$\beta = \text{prob}(\text{Type 2 error})$$
$$= \text{prob}(\text{accepting } H_0/H_a \text{ is False})$$

Ex:

$\mu_1 - \mu_2 = 0$

	Accept H_0	Reject H_a.
H_0 is True.	Correct decision	Type 1 error
H_0 is False	Type 2 error	Correct decision

we aim to reduce both errors, but due to fixed Sample size, It is not possible to ctrl., both the errors Simultaneously.

- To reduce one error we must agree to ↑ the prob., of making the other type of error.

- To reduce β, ↑ α Value.

- Managers decide the significance level.

- It is more dangerous to accept a false hyp., (Type 2 error than to reject a correct one) Type 1 error.

- So we keep the prob of committing Type 1 error at a certain level. Called level of significance.

- Level of significance or size of rejection region or size of critical region or simply size of test

- At 5% level of significance, means that the Prob of accepting a true hypothesis 95%.

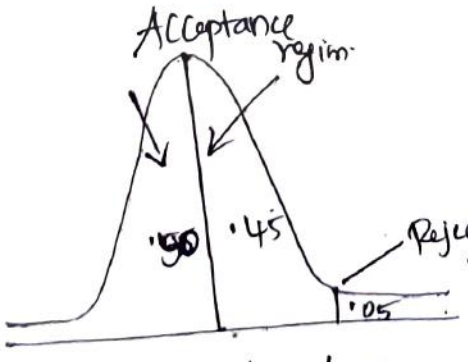
One-tailed test:

It is called so, 'coz the rejection region will be located in one tail which may be either left or right depending upon the alternative hypothesis formulated.

Ex:

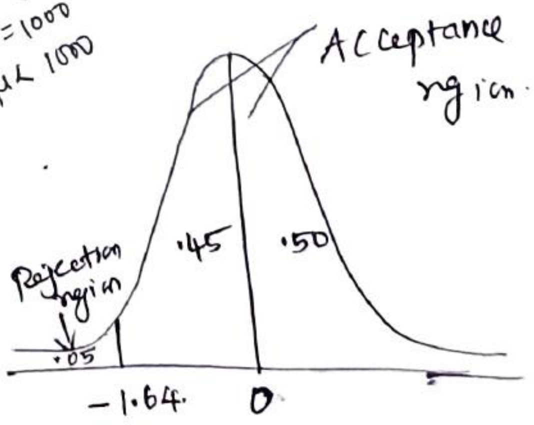
Testing a hypothesis that the average income per household is greater than Rs. 1000, against the alternative hypothesis that the income is Rs. 1000, we will place all the all the alpha risk on right side of the sampling distbn & test will be one-sided right test. $H_0: \mu > 1000$ $H_a: \mu = 1000$

Testing a hypothesis that the average income/household is Rs. 1,000 against alternative that the income is less than Rs. 1000 or less, alpha risk on left side of sampling distbn & the test will be one-sided left tail test.



one-tail $\alpha = 0.05$

$H_0: \mu = 1000$
$H_a: \mu < 1000$



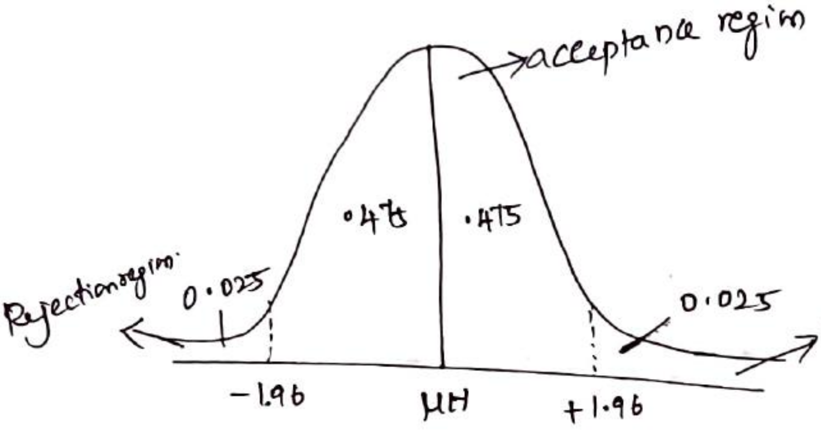
population mean μ_0

$H_0: \mu = \mu_0$

$H_a: \mu \neq \mu_0 (\mu > \mu_0 \text{ or } \mu < \mu_0)$ (2-tailed)

(Right) (Left)
$H_a: \mu > \mu_0$, $H_a: \mu < \mu_0$

Two tailed test :

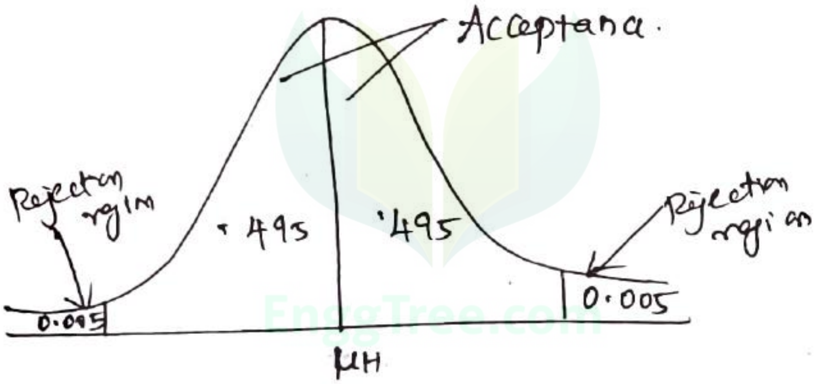


Ho: μ = Mean

H1, μ ≠ Mean

at 0.05

To reduce the risk of committing an error of Type 1. This is done by reducing the size of rejection region.



$\alpha = 0.01$

↓ size of rejection region ↑ prob of accepting our hypothesis

Ex:

The average income / household is Rs. 1000 against the alternative hypothesis that is not Rs. 1000 the rejection region would lie on both sides 'coz we would reject the null hypothesis if the mean income in the sample either too far above Rs. 1000 or too far below Rs. 1000.

20 yrs, is being ~~started~~ with a sample of 40 yr old mean.

$\textcircled{7}$

Soln:: $H_0 : \mu \geq 23$ lbs ; $H_1 : \mu < 23$ lbs

Hint ::

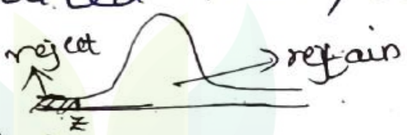
Increase ; upper tail critical

reduce or decrease :- lower tail critical

pblm ::

For each situation state whether Ho rejected or retained.

a) Given a one-tailed test, lower tail critical with $\alpha = .01$ and



a) $z = -2.34$

$z_c = -2.33$

Ho : reject

b) $z = -5.13$

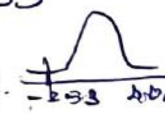
$z_c = -2.33$

Ho : reject

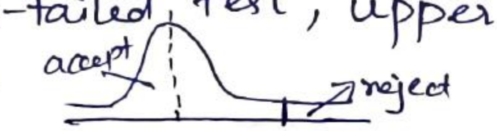
c) $z = 4.04$

$z_c = -2.33$

Ho : retain



b) Given a one-tailed test, upper tail critical with $\alpha = .05$ &



d) $z = 2.00$

$z_c = 1.65$

Ho : reject

e) $z = -1.80$.

$z_c = +1.65$

Ho : retain

f) $z = 1.61$

$z_c = +1.65$

Ho : retain

For one tailed tests, the new null hypothesis ·

one-tailed @ lower-tail critical

$H_1: \mu < 500$ then $H_0: \mu \geq 500$ Instead of $H_0: \mu = 500$.

@ upper tail critical.

$H_1: \mu > 500$, then $H_0: \mu \leq 500$.

Ex:

a) An investigator wishes to determine whether, for a sample of drug addicts, the mean score on the depression scale of a personality test differs from a score of 60, is mean score of general population

Soln:

$H_0: \mu = 60$; $H_1: \mu \neq 60$

b) To ↑ rainfall, extensive cloud-seeding expts are to be conducted & results are to compared with a baseline fig of 0.54 inch of rainfall.

$H_0: \mu \leq 0.54$; $H_1: \mu > 0.54$.

c) Public health statistics indicate, we will assume, that american males gain an avg of 23 lbs during 20yr period after 40yrs of age. An ambitious weight-reduction pgm spanning

Rejecting Ho is a strong Decision:

Ho is rejected whenever the Zo qualifies as rare outcome - with prob of .05 or less on assumption that Ho is true.

Estimation:

Statistical estimation or estimation is concerned with the methods by which population characteristics are estimated from sample infmn..

→ Finite populations - cost of complete censuses may be prohibitive.

→ Infinite population - complete enumerations are impossible.

→ desired degrees of populations

2 types

1) point estimates

2) Interval estimates

Testing a hypothesis about Vitamin c:

- To determine whether Vitamin C increases the intelluctual aptitude of high school students. After being randomly selected from some large school district, each of 36 students takes a daily dose of 90 mg of Vit C for a period of 2 months b4 being tested for IQ.

Ordinarily, IQ scores for all students in this school district approximate a nml., distbn with a mean of 100 and a std deviation of 15.

Research Pblm:

Does the daily ingestion of Vit C cause an increase, on average in IQ scores among all students in the school dt?

Hypothesis: upper tail critical

$H_0: \mu \leq 100$; $H_1: \mu > 100$

Decision rule:

Reject Ho at the 0.05 level of significance if $Z \geq 1.65$

Calculations:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{36}} = 2.5;$$

$$Z_0 = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{100-100}{25} = 0.4$$

Critical Z-Values.

Type of test	Level of Significance (α)	
	.05	.01.
Two-tailed or non-directional ($H_0 : \mu =$ Some no) ($H_1 : \mu \neq$ Some no)	± 1.96	± 2.58
one-tailed or directional test, lower tail critical. $H_0 : \mu \geq$ Some no $H_1 : \mu <$ Some no	-1.65	-2.33
One-tailed or directional test, upper tail critical $H_0 : \mu \leq$ some no $H_1 : \mu >$ some no	$+1.65$	$+2.33$

* Specify the decision rule for each of the foll., situations

a) two-~~test~~tailed test with $\alpha = -0.05$, ± 1.96

Reject H_0 @ 0.05 level of Significance if 'Z' equals or is more positive than 1.96 or Z equals or is more -ve than -1.96.

b) a one tailed test, upper tail critical with $\alpha = .01$

c) a one tailed test, lower tail critical with $\alpha = .05$

d) a two tailed test with $\alpha = .01$

	Accept	[illegible]
Ho True	CD	T-I
Ho False	T-II	CD

Z-test is approoriate:

Sample = 36 $n > 30$

4 possible outcomes!

i) If Ho is really true. (i.e. Vit C does not cause an ↑ in the population mean IQ), then it is a correct decision to retain the true Ho. In this case, we could conclude correctly that there is no evidence that Vit C ↑ IQ.

ii) If Ho is really true, it is type-I error to reject the true Ho and conclude that Vitamin C ↑ IQ when it doesn't do. This will allow us to chare after something does not exist.

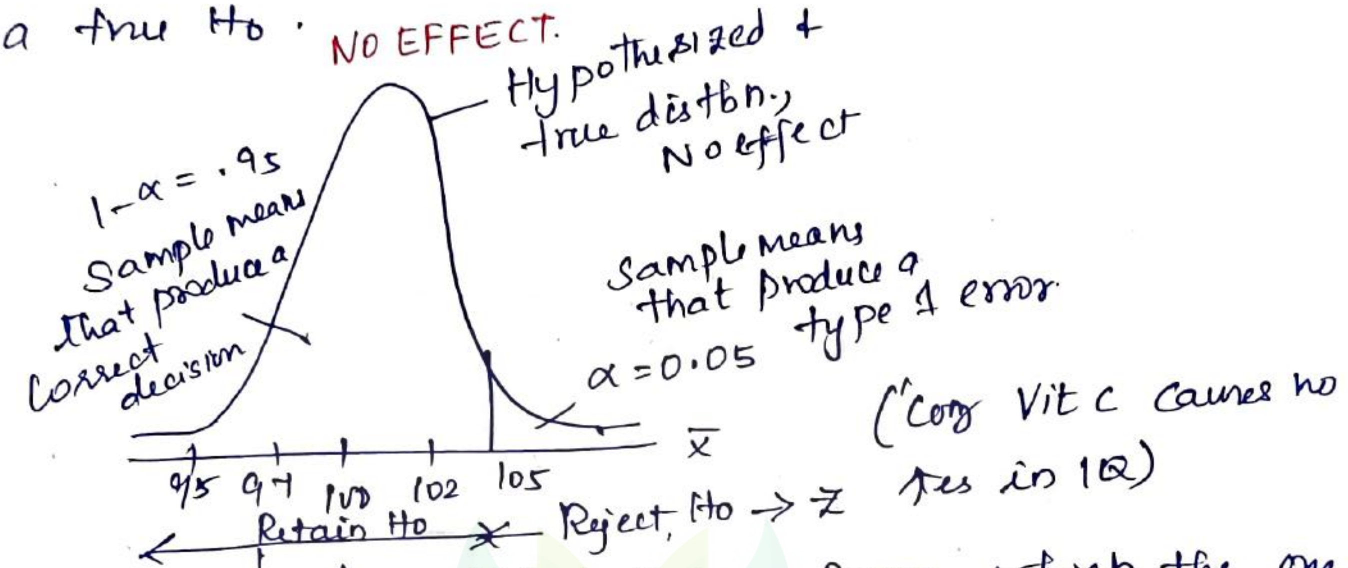
iii) If Ho is really false (i.e. Vit C really causes an ↑ in the population mean IQ), then it is a type-II error to retain the false Ho and conclude that there is no evidence that Vit C increases IQ, when infact it does. It is called misses, coz they fail to detect a potentially important relationship, such as that b/w Vit C & IQ.

iv) If Ho really is false, then it is a correct decision to reject the false Ho & conclude that Vit C ↑ IQ.

If Ho Really is true.

'Coz Vit C' doesn't ↑ the population mean IQ.

In this case, we concern about retaining or rejecting a true Ho.


NO EFFECT.
Hypothesized + true distbn., No effect
$1-\alpha = .95$
Sample means that produce a correct decision
Sample means that produce a type I error.
$\alpha = 0.05$
$\bar{x}$
('Coz Vit C causes no ↑es in IQ)
95
97
100
102
105
Retain Ho
Reject Ho → Z

- True sampling distribution - from which the one observed sample mean actually Originates.

∴, the one observed sample mean in the expt can be viewed as being random selected from the the hypothesized distribution.

- If Ho is really true, the proba of type I error, α equals the level of Significance and the probability of a correct decisions equals $1-\alpha$.

- The level of Significance indicates the proport Of the total area of the sampling distribution in the rejection region for Ho.

Reducing the prob of T-I error:

- If Ho is really true, the present test will produce a correct decision with a prob of .95 & Type I error with a prob of .05.

- If a false alarm has serious consequences, the prob of a type I error can be reduced to .01 or even to .001 simply by using .01 or .001 level of significance.

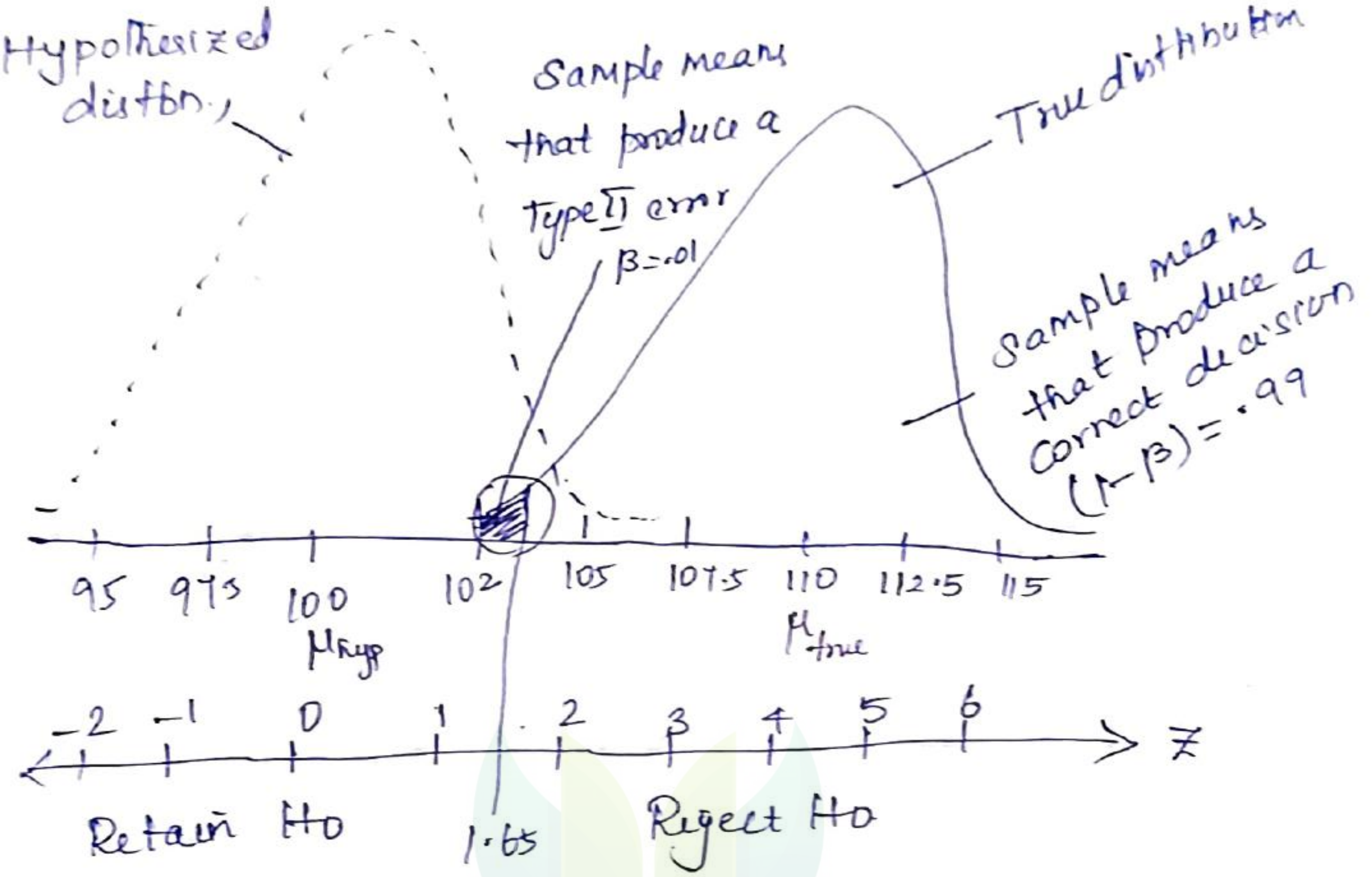
- Any one of the level of significance is preferred, for the Vit C test if, for instance, a false alarm could cause the adoption of an expensive pgm to supply worthless Vit C to all students in the district and, perhaps, the creation of an accelerated curriculum to accomodate the fictitious increase in intellectual aptitude.

If Ho Really is False, because of a large effect:

If Ho is really false / coz Vit C ↑es the population mean by not just few pts, but by many points for ex 10 pts.

→ Any difference b/w true & hypothesized ~~sampling distbn~~ population mean is effect:

Large effect.



- It is essential to distinguish b/w hypothesized sampling diston and the true sampling diston.
- Centered about the hypothesized population mean of 100, it serves as a parent diston for the familiar decision rule with a critical z of 1.65 for the projected one-tailed test.

True Sampling distribution:

If mean is ↑ by 10 points then true sampling diston., serves as the parent distribution for the one randomly selected sample mean that will be observed in the expt.

- Viewed relative to the decision rule, the one randomly selected sample mean dictates whether we retain or reject the false Ho.

Low probability of a Type II error for a Large effect:

- A randomly selected sample mean Originates from the very small black portion of the true sampling distbn. of the mean, its 'Z' value is less than 1.65 and ∴ Ho is retained.
- If ∴ Ho is really false, this is an incorrect decision or type-II error - a miss announced as a lack of evidence that Vit C ↑ IQ, even though it does.
- The prob of a type-II error, Symbolized by the letter (β) equals 0.01

High prob of a correct Decision for a large effect.

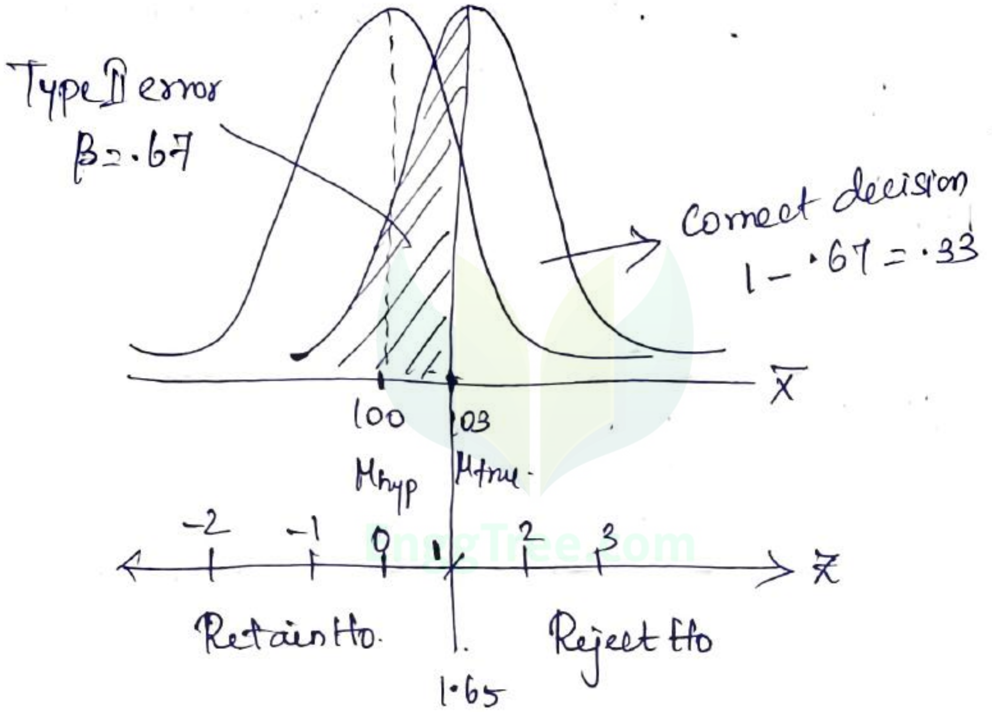
- When a Sample mean Originates from the large shaded portion of the true sampling distbn, its Z value equals or exceeds 1.65 and Ho is rejected.
- Ho really is false, then it is a Correct decision announced as evidence that Vit 'C' increases IQ.
- Then the prob of Correct decision $1-\beta = .99$

→ One sided test performs well.

If Ho really is False, because of a small effect.

- Case study for Vit C increases the pop, mean IQ by only a few points - 3 pts.
- Then Hypothesis Sampling Distbn - Centered at 100

True sampling distbn - Centered at 103.



* Low prob of a correct decision for a small effect.

→ W.r.t the decision rule, the true sampling distbn supplies two types of randomly selected sample means:

i) that produce type II error, coz they Originate from shaded Region.

ii) those that produce a correct decision coz they Originate from the white part.

- Because of Smaller effect, the true and hypothesized population means are much closer.

- As a result, the entire true sampling distbn, is shifted towards the retention region for the false Ho.

- Now the projected one-tailed test performs more poorly; there is high probability of 0.67 that a type II error will be committed & low-probability 0.33 that correct decision will be made.

(7) If Ho is really false, the probability of a type-II error, β and the prob of correct decision (1-β) depend on the size of the effect, that is diff, b/w the true and the hypothesized population means. The smaller the effect, the higher the probability of a type II error and the lower the probability of a correct decision.

Influence of Sample size:

→ w.r.t the previous ex. the investigator might not be too concerned about the low detection rate of 0.33 for the relatively small three-point effect of Vit C on IQ.

→ But Special cases make us worry about it.

→ for the previow expt, Vit C has many positive effects including the reduction in the duration and severity of common colds and no side effects. furthermore, huge quantities of Vit C might be available at no cost to the school district.

→ Another point, a fairly mild one such as a small increase in the pop., mean IQ, might clinch the supply of Vit C. to all the students in the district.

• To increase the prob of detecting a false Ho, ↑ the sample size.

→ Assume still the expt has 3-point effect on IQ, we can check the properties of the projected one-tailed test when the sample size is increased from 36 to 100 students.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

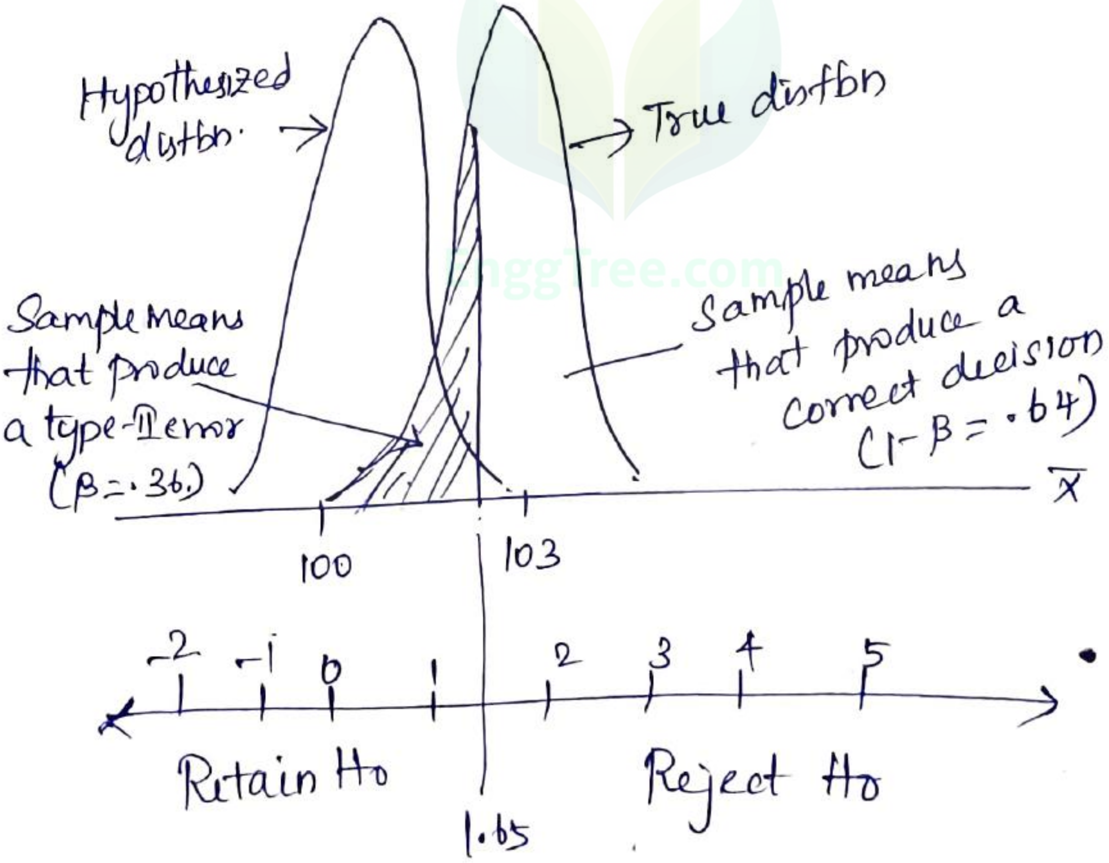
for the Original experiment with its sample size of 36,

$$\sigma_{\bar{x}} = \frac{15}{\sqrt{36}} = \frac{15}{6} = 2.5$$

for $n = 100$

$$\sigma_{\bar{x}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1.5$$

When 'n' ↑ '$\sigma_{\bar{x}}$' ↓.


Hypothesized distbn.
True distbn
Sample means that produce a type-II error ($\beta = .36$)
Sample means that produce a correct decision ($1-\beta = .64$)
$\bar{X}$
100
103
-2
-1
0
1
2
3
4
5
Retain Ho
Reject Ho
1.65

~~Draw back~~

Drawbacks of reducing $\sigma_{\bar{X}}$.

1) It shrinks the upper retention region back toward the hypothesized population mean of 100.

2) It shrinks the entire true sampling distbn toward the true population mean of 103.

From the figure, it is understood for 100 students, fewer sample means (.36) produce a type II error 'cony they originate from the shaded portion & correct decision (.64) lead to detection of Ho.

→ If $\sigma_{\bar{X}}$ is further reduced, the upper retention region shrinks to the immediate vicinity of the hypothesized population mean of 100 and the entire true sampling distbn., of the mean, shrinks to the immediate vicinity of the true population mean of 103. The net result is that type II error hardly ever is committed and the small 3-pt effect virtually is detected.

power and sample size :

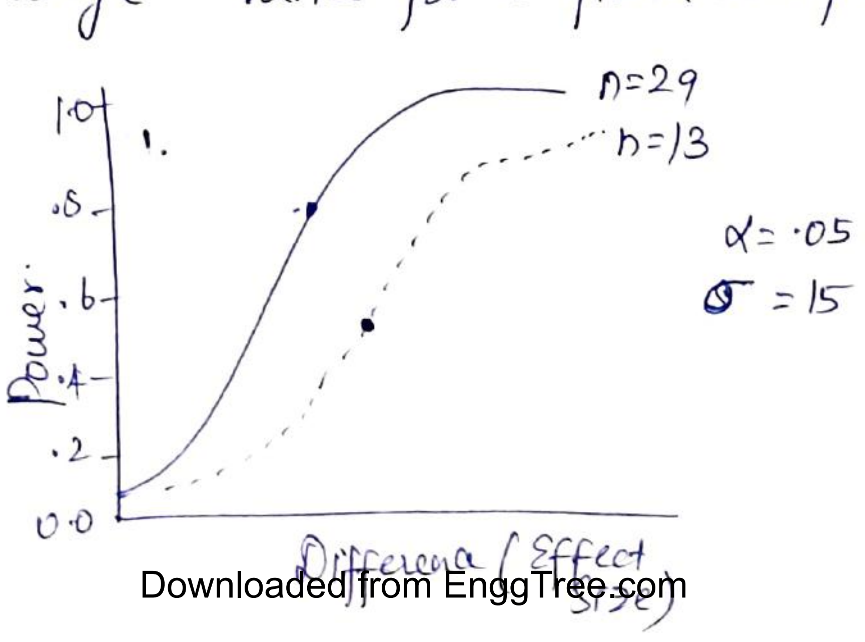
The power of a hypothesis test equals the prob (1-β) Of detecting a particular effect when the null hypothesis is False.

- power is simply the Complement of (1-β) of the prob., 'β' of failing to detect the effect.

- Selection of sample size Should reflect
  i) the smallest important effect.
  ii) reasonable degree of power for detecting that effect.

Power curves:

- It shows ~~the~~ how the likelihood of detecting any possible effect - ranging from very small to very large - Varies for a fixed sample size.



The dot on the power curve for a sample of 13 indicates that 7-pt effect will be detected with the power of approximately 0.50.

Point Estimate for μ:

A point estimate for μ uses a single value to represent the unknown population mean.

For SAT score example,

$\mu = 533$

↳ point estimate.

→ Drawback is inaccurate.

→ So we use alternative interval estimates or confidence interval.

Confidence interval (CI) for μ.

CI for 'μ' uses a range of values that with a known degree of certainity, includes the unknown population mean.

Why Confidence. Intervals Work?

3 important properties

i) The mean of the SD equals the unknown population mean for all local freshman.

ii) The Std, error of the sampling distribution equals the value

iii) The shape of the sampling distbn approximates a normal distribution because the sample size of 100 satisfies the reqts, of central limit theorem.

Series of confidence Interval:

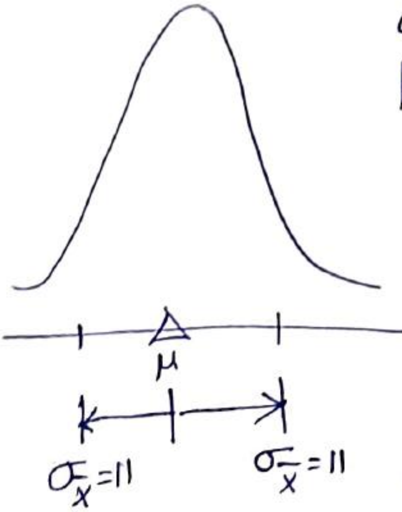
→ practically, only one sample mean is actually taken from sampling distbn. and used to construct a single 95% CI.

→ Imagine a series of samples from sampling distbn, For each sample mean, construct a 95% CI by adding $1.96\sigma_{\bar{x}}$ to the sample mean & subtracting $1.96\sigma_{\bar{x}}$ from the sample mean.

$$\bar{X} \pm 1.96\sigma_{\bar{X}}$$

to obtain a 95% confidence interval for each sample mean.

True Confidence Interval:

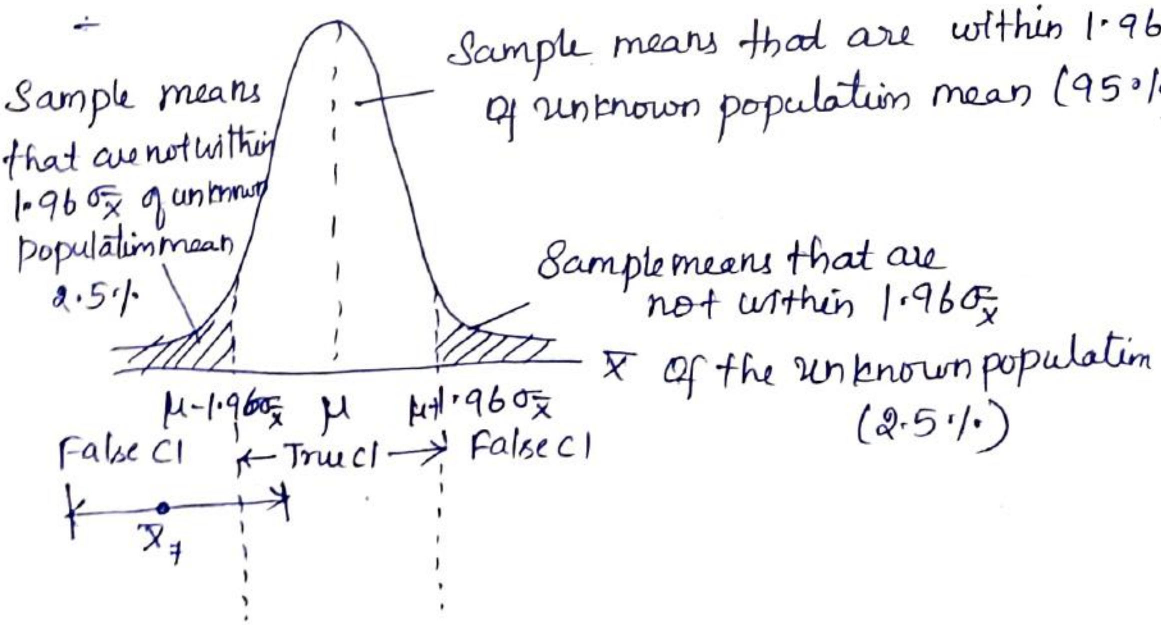


do 95% of these CI include the unknown population mean?

From the figure, Sampling distbn, is normal 95% of all sample means are within 1.96 Std., errors of the unknown population mean (i.e) 95% of all sample means deviate less than 1.96 Standard errors from the unknown population mean.

False Confidence interval:

- 5% of all confidence intervals fail to include the unknown population mean.


Sample means that are within $1.96\sigma_{\bar{x}}$ of unknown population mean (95%)
Sample means that are not within $1.96\sigma_{\bar{x}}$ of unknown population mean 2.5%
Sample means that are not within $1.96\sigma_{\bar{x}}$
$\bar{x}$ of the unknown population (2.5%)
$\mu-1.96\sigma_{\bar{x}}$
μ
$\mu+1.96\sigma_{\bar{x}}$
False CI
True CI
False CI
$\bar{x}_7$

CI for μ based on z.

$$\bar{x} \pm (z_{conf})(\sigma_{\bar{x}})$$

Given $\bar{x} = 533$, $z_{conf} = 1.96$, $\sigma_{\bar{x}} = 11$

$$533 \pm (1.96)(11) = 533 \pm 21.56 = \begin{cases} 554.56 \\ 511.44 \end{cases}$$

When the level of confidence equals 95% or more we can be reasonably confident that the one observed confidence interval include the true population mean

- The level of confidence indicates the percent of time that a series of CI includes the unknown population characteristic such as the population mean.

For 99% $z_{conf} = \pm 2.58$

$$533 \pm (2.58)(11) = 533 \pm 28.38 = \begin{cases} 561.38 \\ 504.62 \end{cases}$$

Effect on width of interval:

- at 99% CI of 504.62 to 561.38 is wider and ∴ less precise than the corresponding 95% confidence interval of 511.44 to 554.56.
- The shift from a 95% to a 99% level of confidence requires an increase in the value of z_{conf} from 1.96 to 2.58. This increase causes a wider, less precise confidence interval.

Choosing a level of Confidence:

→ Although many different levels of confidence have been used, 95% and 99% are the most prevalent.

→ a larger level of Confidence, such as 99% Should be reserved for situations in which a false interval might have particularly serious consequences.

Effect of Sample Size:

→ The larger the sample size, the smaller the $\sigma_{\bar{x}}$ and hence the more precise the Confidence interval will be.

Hypothesis Tests or Confidence intervals?

It indicates whether or not an effect is present, whereas confidence intervals indicate the possible size of the effect.

Confidence interval for population percent:

Sample percent ± (1.96) (Std., error of the percent)

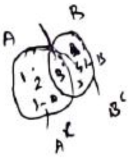
Sample size and Margin of Error:

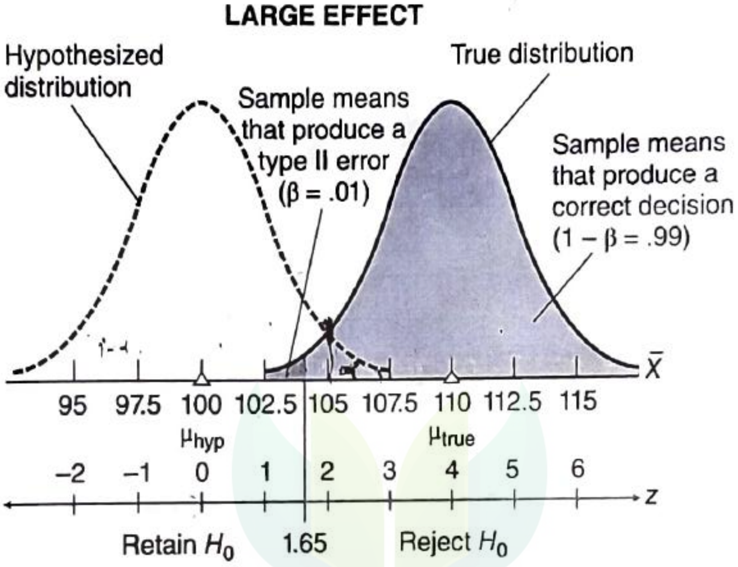
If 'n' ~~is~~ is high, Margin of Error is less.

for 1500, ±3 margin of Error.
500, ±5 "
100, ±10 "

UNIT II -PROBLEMS

1. Indicate whether the following statements, all referring to Figure 11.4, are true or false: (a) The assumption that H0 really is false is depicted by the separation of the hypothesized and true distributions. (b) In practice, when actually testing a hypothesis, we would not know that the true population mean equals 110. (c) The one observed sample mean is viewed as originating from the hypothesized sampling distribution. (d) A correct decision would be made if the one observed sample mean has a value of 103.





Solution:

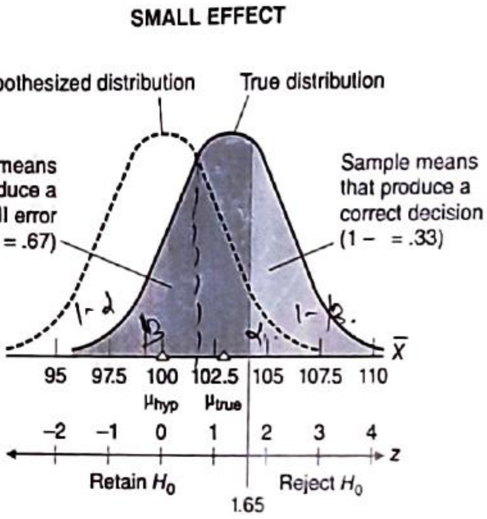
(a) True

(b) True

(c) False. The one observed sample mean originates from the true sampling distribution.

(d) False. If the one observed sample mean has a value of 103, an incorrect decision would be made because the false H0 would be retained

2. Indicate whether the following statements, all referring to Figure 11.5, are true or false: (a) The value of the true population mean (103) dictates the location of the true sampling distribution. (b) The critical value of z (1.65) is based on the true sampling distribution. (c) Since the hypothesized population mean of 100 really is false, it would be impossible to observe a sample mean value less than or equal to 100. (d) A correct decision would be made if the one observed sample mean has a value of 105.



Solution:

a) True

(b) False. The critical value of z (1.65) is based on the hypothesized sampling distribution.

(c) False. Since the true sampling distribution goes below 100, a sample mean less than or equal to 100 is possible, although not highly likely.

(d) True

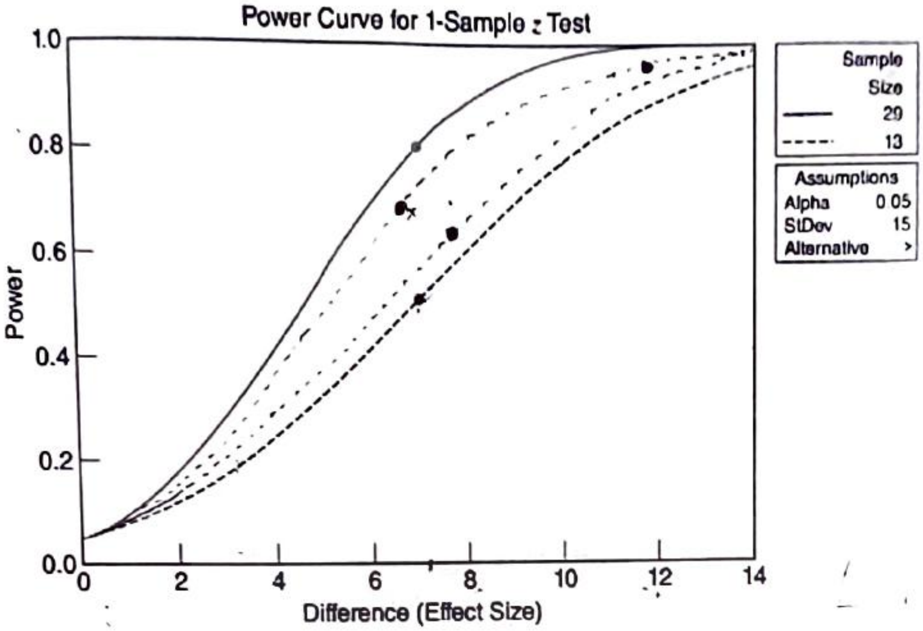
3. Comment critically on the following experimental reports: (a) Using a group of 4 subjects, an investigator announces that H0 was retained at the .05 level of significance. (b) Using a group of 600 subjects, an investigator reports that H0 was rejected at the .05 level of significance.

Solution:

(a) Because of the small sample size, only very large effects will be detected.

(b) Because of the large sample size, even small, unimportant effects will be detected.

4. Consult the power curves in Figure 11.7 to estimate the approximate detection rates, rounded to the nearest tenth, for the following situations: (a) a three-point effect, with a sample size of 29 (b) a six-point effect, with a sample size of 13 (c) a twelve-point effect, with a sample size of 13.



Solution:

(a) .3

(b) .4

(c) .9

5. An investigator consults a chart to determine the sample size required to detect an eight-point effect with a probability of .80. What happens to this detection rate of .80—will it actually be smaller, the same, or larger—if, unknown to the investigator, the true effect actually equals (a) twelve points? (b) five points?

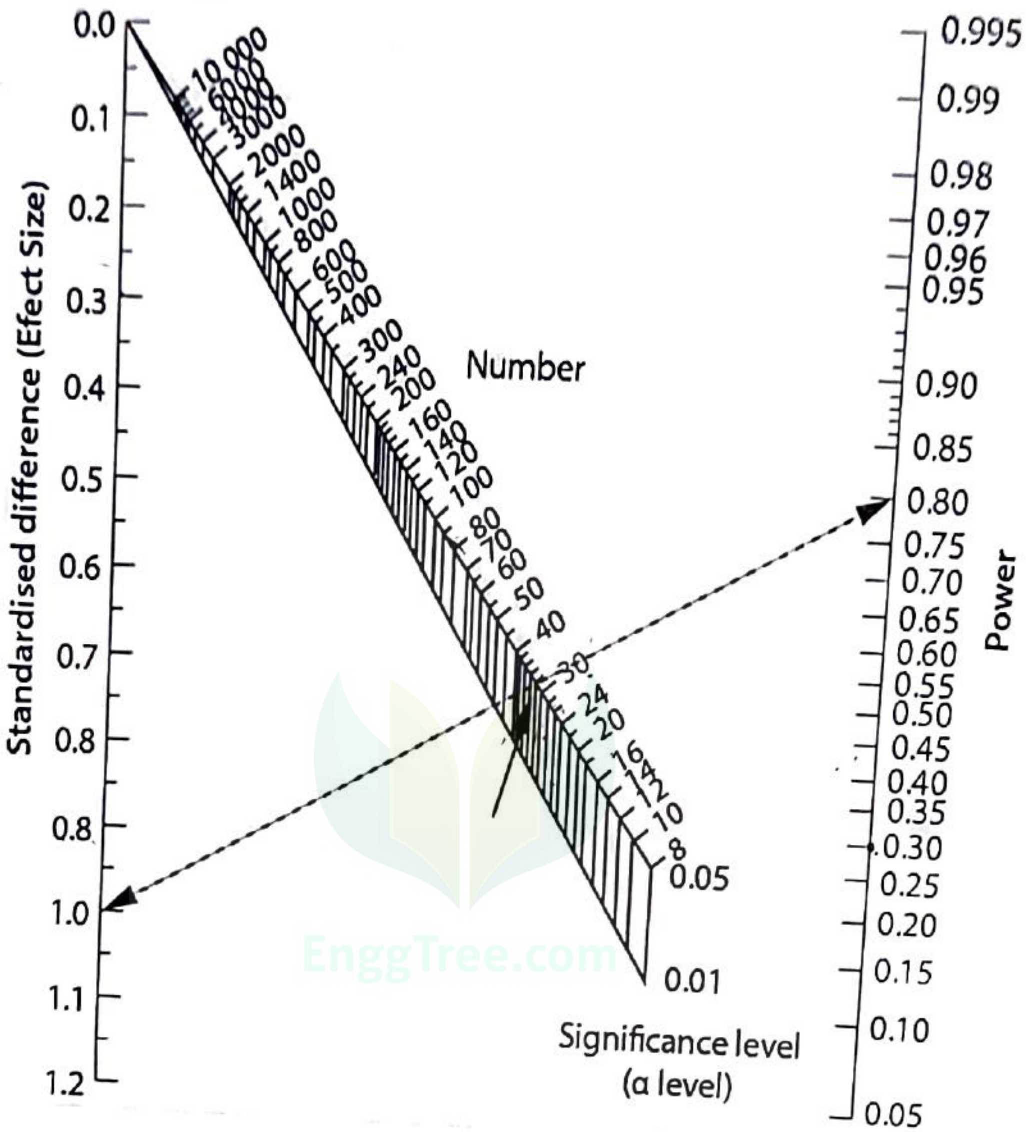
Effect ↑ power ↑.

Solution:

a) The power for the 12-point effect is larger than .80 because the true sampling distribution is shifted further into the rejection region for the false H0.

(b) The power for the 5-point effect is smaller than .80 because the true sampling distribution is shifted further into the retention region for the false H

accept


Standardised difference (Efect Size)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.8
1.0
1.1
1.2
Number
10,000
6000
4000
3000
2000
1400
1000
800
600
500
400
300
240
200
160
140
120
100
80
70
60
50
40
30
24
20
16
14
12
10
8
0.05
0.01
Significance level
(α level)
Power
0.995
0.99
0.98
0.97
0.96
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05

***11.14** Recalling the vitamin C experiment described in this chapter, you could describe the null hypothesis in both symbols and words as follows:

$H_0 : \mu \leq 100$, that is, vitamin C does not increase IQ

Following the format of Table 11.2 and being as specific as possible, you could describe the four possible outcomes of the vitamin C experiment as follows:

	STATUS OF H_0	
DECISION	**TRUE H_0**	**FALSE H_0**
Retain H_0	*Correct Decision:* Conclude that there is no evidence that vitamin C increases IQ when in fact it doesn't.	*Type II Error:* Conclude that there is no evidence that vitamin C increases IQ when in fact it does.
Reject H_0	*Type I Error:* Conclude that vitamin C increases IQ when in fact it doesn't.	*Correct Decision:* Conclude that vitamin C increases IQ when in fact it does.

Using the answer for the vitamin C experiment as a model, specify the null hypothesis and the four possible outcomes for each of the following exercises:

To increase rainfall, with baseline $\underset{H_0}{\mu \leq 0.5}$

11.14 (a) $H_0: \mu \leq 0.54$, that is, cloud seeding has no effect on rainfall.

	STATUS OF H_0	
DECISION	**TRUE H_0**	**FALSE H_0**
Retain H_0	*Correct Decision:* Conclude that there is no evidence that cloud seeding increases rainfall when in fact it does not.	*Type II Error:* Conclude that there is no evidence that cloud seeding increases rainfall when in fact it does.
Reject H_0	*Type I Error:* Conclude that cloud seeding increases rainfall when in fact it does not.	*Correct Decision:* Conclude that cloud seeding increases rainfall when infact it does.

6. A production line at a candy plant is designed to yield 2-pound boxes of assorted candies whose weights in fact follow a normal distribution with a mean of 33 ounces and a standard deviation of .30 ounce. A random sample of 36 boxes from the production of the most recent shift reveals a mean weight of 33.09 ounces. (Incidentally, if you think about it, this is an exception to the usual situation where the investigator hopes to reject the null hypothesis.)

(a) Describe the population being tested.

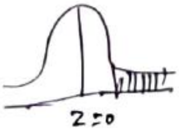
(b) Using the customary procedure, test the null hypothesis at the .05 level of significance.

$H_0 : \mu = 33$
$H_1 : \mu \neq 33$

(c) Someone uses a one-tailed test, upper tail critical, because the sample mean of 33.09 exceeds the hypothesized population mean of 33. Any comment?

$\mu = 33; \ n = 36; \ \bar{x} = 33.09$

Solution: $H_0 : \mu \leq 33$, $H_1 : \mu > 33$



$z = 33.09 -$

7. Consult the power curves in Figure 11.7 to estimate the approximate detection rate, rounded to the nearest tenth, for each of the following situations: (a) a four-point effect, with a sample size of 13 (b) a ten-point effect, with a sample size of 29 (c) a seven-point effect with a sample size of 18 (Interpolate).

Solution:

8. A random sample of 200 graduates of U.S. colleges reveals a mean annual income of $62,600. What is the best estimate of the unknown mean annual income for all graduates of U.S. colleges?

Solution:

$62,600

9. Reading achievement scores are obtained for a group of fourth graders. A score of 4.0 indicates a level of achievement appropriate for fourth grade, a score below 4.0 indicates underachievement, and a score above 4.0 indicates overachievement. Assume that the population standard deviation equals 0.4. A random sample of 64 fourth graders reveals a mean achievement score of 3.82. (a) Construct a 95 percent confidence interval for the unknown population mean. (Remember to convert the standard deviation to a standard error.) (b) Interpret this confidence interval; that is, do you find any consistent evidence either of overachievement or of underachievement?

Solution:

(a) $3.82 \pm 1.96 .4\ 64 = 3.92\ 3.72$

(b) We can claim, with 95 percent confidence, that the interval between 3.72 and 3.92 includes the true population mean reading score for the fourth graders. All of these values suggest that, on average, the fourth graders are underachieving.

10. Before taking the GRE, a random sample of college seniors received special training on how to take the test. After analyzing their scores on the GRE, the investigator reported a dramatic gain, relative to the national average of 500, as indicated by a 95 percent confidence interval of 507 to 527. Are the following interpretations true or false? (a) About 95 percent of all subjects scored between 507 and 527. (b) The interval from 507 to 527 refers to possible values of the population mean for all students who undergo special training. (c) The true population mean definitely is between 507 and 527. (d) This particular interval describes the population mean about 95 percent of the time. (e) In practice, we never really know whether the interval from 507 to 527 is true or false. (f) We can be reasonably confident that the population mean is between 507 and 527.

Solution:

(a) False. We can be 95 percent confident that the mean for all subjects will be between 507 and 527.

(b) True

(c) False. We can be reasonably confident—but not absolutely confident—that the true population mean lies between 507 and 527.

(d) False. This particular interval either describes the one true population mean or fails to describe the one true population mean.

(e) True

(f) True

11. On the basis of a random sample of 120 adults, a pollster reports, with 95 percent confidence, that between 58 and 72 percent of all Americans believe in life after death. (a) If this interval is too wide, what, if anything, can be done with the existing data to obtain a narrower confidence interval? (b) What can be done to obtain a narrower 95 percent confidence interval if another similar investigation is being planned?

Solution:

(a) Switch to an interval having a lesser degree of confidence, such as 90 percent or 75 percent.

(b) Increase the sample size.

12. In a recent scientific sample of about 900 adult Americans, 70 percent favor stricter gun control of assault weapons, with a margin of error of ±4 percent for a 95 percent confidence interval. Therefore, the 95 percent confidence interval equals 66 to 74 percent. Indicate whether the following interpretations are true or false: (a) The interval from 66 to 74 percent refers to possible values of the sample percent. (b) The true population percent is between 66 and 74 percent. (c) In the long run, a series of intervals similar to this one would fail to include the population percent about 5 percent of the time. (d) We can be reasonably confident that the population percent is between 66 and 74 percent.

Solution:

(a) False. The interval from 66 to 74 percent refers to possible values of the population proportion.

(b) False. We can be reasonably confident—but not absolutely confident—that the true population proportion is between 66 and 74 percent.

(c) True

(d) True

13. In Question 10.5 on page 191, it was concluded that, the mean salary among the population of female members of the American Psychological Association is less than that ($82,500) for all comparable members who have a doctorate and teach full time. (a) Given a population standard deviation of $6,000 and a sample mean salary of $80,100 for a random sample of 100 female members, construct a 99 percent confidence interval for the mean salary for all female members. (b) Given this confidence interval, is there any consistent evidence that the mean salary for all female members falls below $82,500, the mean salary for all members?

Solution:

(a) 80,100 ± 2.58 6,000 100 = 81,648 78,552

(b) We can claim, with 99 percent confidence, that the interval between $78,552 and $81,648 includes the true population mean salary for all female members of the American Psychological Association. All of these values suggest that, on average, females' salaries are less than males' salaries.

14. Imagine that one of the following 95 percent confidence intervals estimates the effect of vitamin C on IQ scores:

95% CONFIDENCE INTERVAL	LOWER LIMIT	UPPER LIMIT
1	100	102
2	95	99
3	102	106
4	90	111
5	91	98

(a) Which one most strongly supports the conclusion that vitamin C increases IQ scores?

(b) Which one implies the largest sample size?

(c) Which one most strongly supports the conclusion that vitamin C decreases IQ scores?

(d) Which one would most likely stimulate the investigator to conduct an additional experiment using larger sample sizes?

Solution:

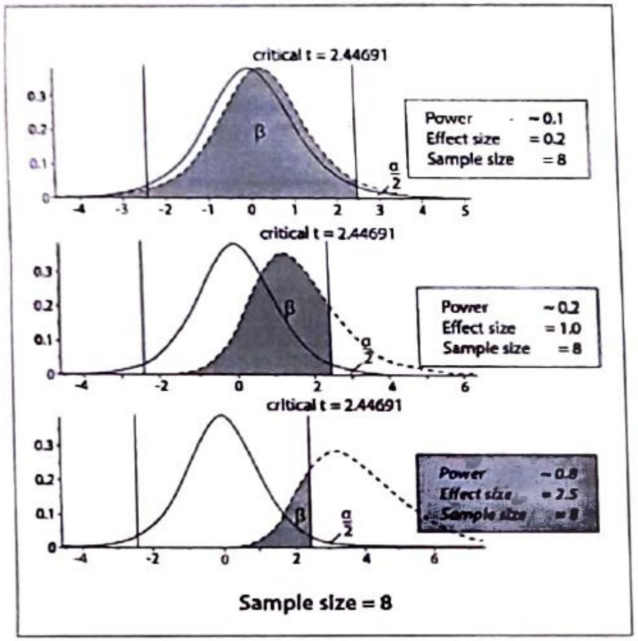
(a) 3

(b) 1

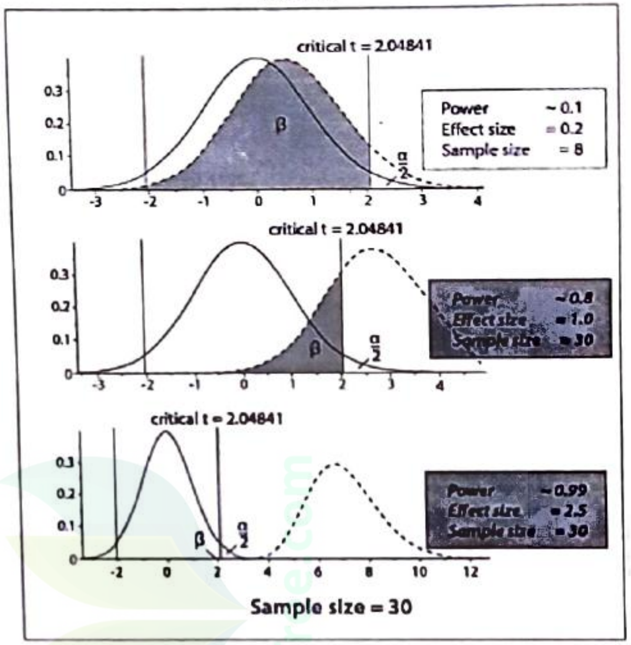
(c) 5

(d) 4

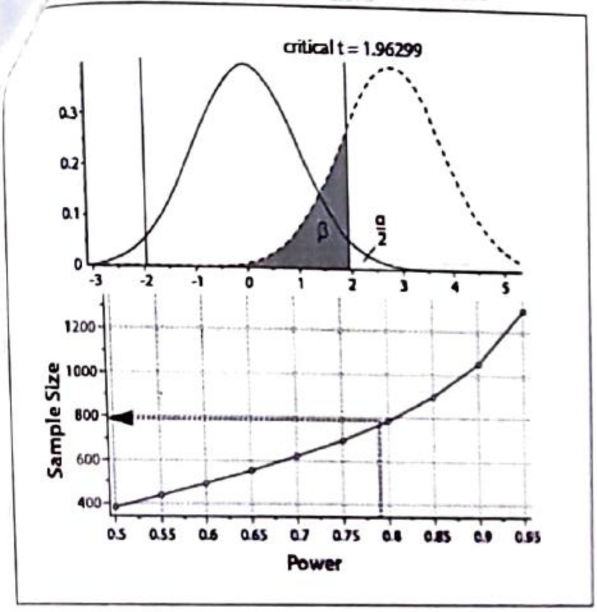
Case 4


critical t = 2.44691
Power ~ 0.1
Effect size = 0.2
Sample size = 8
β
$\frac{\alpha}{2}$
0.3
0.2
0.1
0
-4 -3 -2 -1 0 1 2 3 4 5
critical t = 2.44691
Power ~ 0.2
Effect size = 1.0
Sample size = 8
-4 -2 0 2 4 6
critical t = 2.44691
Power ~ 0.8
Effect size = 2.5
Sample size = 8
-4 -2 0 2 4 6
Sample size = 8

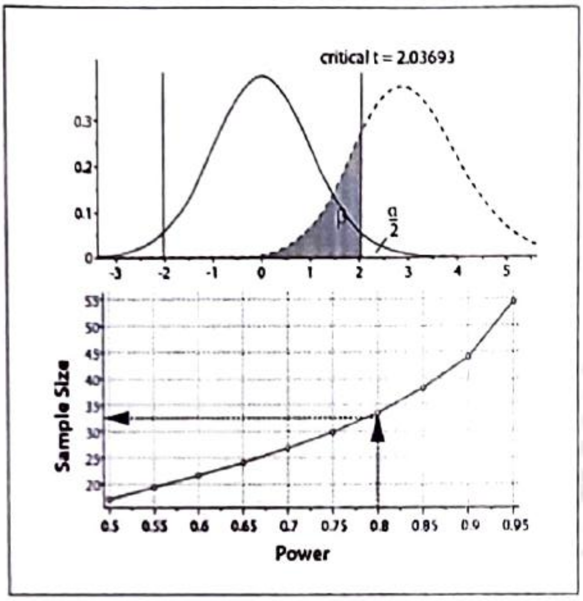
Case 5


critical t = 2.04841
Power ~ 0.1
Effect size = 0.2
Sample size = 8
β
$\frac{\alpha}{2}$
0.3
0.2
0.1
0
-3 -2 -1 0 1 2 3 4
critical t = 2.04841
Power ~ 0.8
Effect size = 1.0
Sample size = 30
-3 -2 -1 0 1 2 3 4
critical t = 2.04841
Power ~ 0.99
Effect size = 2.5
Sample size = 30
-2 0 2 4 6 8 10 12
Sample size = 30

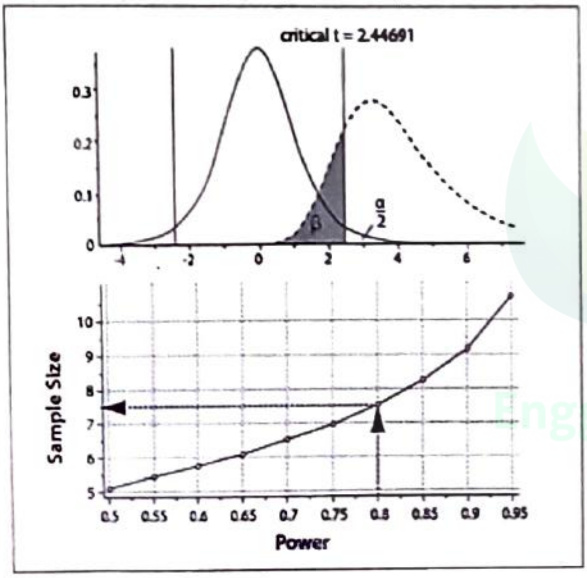
Case 1: *P = 0.8*, ES = 0.2, SS = 788



Case 2: *P = 0.8*, ES = 1, SS = 34



Case 3: *P = 0.8*, ES = 2.5, SS = 8



Unit-II

Inferential Statistics-II

Why Hypothesis Test?

The variability among sample means must be considered when we attempt to decide whether the observed difference b/w 533 and 500 is real or merely transitory.

Importance of standard error:-

When expressed as z, the ratio of the observed diff., to the standard error is small enough to be reasonably attributed to chance, we retain Ho. otherwise, if the ratio of the observed difference to the standard error is too large to be reasonably attributed to chance, as in the SAT example, we reject Ho.

Before generalizing beyond the existing data, we must measure the effect of chance; that we should obtain a value for the std., error.

for SAT example, increase $\sigma_{\bar{x}}$ from 11 to 33 and note that even though the observed difference remains the same (533-500), we would retain Ho b'coz now z' would equal '1'.

$$z_0 = \frac{533-500}{33} = \frac{33}{33} = 1 \; ; \; z_c = 1.96$$

$z_0 < z_c$ ⇒ condition satisfied so we retain Ho

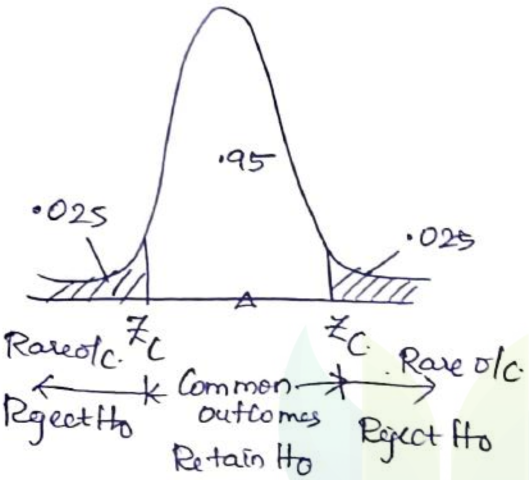
Possibility of Incorrect Decisions

Type-I error or false alarm.

- rejecting a true H_0.

Type-II error or Miss

- retaining a false H_0



Strong or Weak Decisions:

Retaining Ho is a weak Decision

- It is difficult to retain Ho & reject Ho.

→ Ho is retained if 'Z_0' qualifies as common outcome. & so Ho retained.

→ However, the same observed result also would qualify as a common outcome when the Original value Ho(500) is replaced witha slightly different value. If so retaining Ho is weak decision.

→ Statisticians describe it as failure to reject Ho rather than retention of Ho.

Two types of Errors in hypothesis Testing :

Four possibilities.

i) The hypothesis is true but our test rejects it. (Type 1 error)

ii) The hypothesis is false but our test accepts it. (Type 2 error)

iii) Hyp - True, acept test

iv) Hyp - false, reject test.

Type 1 error - represented by 'α'

$$\alpha = \text{Prob}(\text{Type 1 error})$$
$$= \text{Prob}(\text{Rejecting } H_0/H_a \text{ is True})$$

Type 2 error - represented by 'β'

$$\beta = \text{prob}(\text{Type 2 error})$$
$$= \text{prob}(\text{accepting } H_0/H_a \text{ is False})$$

Ex:

$\mu_1 - \mu_2 = 0$

	Accept Ho	Reject Ha.
Ho is True.	Correct decision	Type 1 error
Ho is False	Type 2 error	Correct decision

we aim to reduce both errors, but due to fixed Sample size, It is not possible to ctrl., both the errors Simultaneously.

- To reduce one error we must agree to ↑ the prob., of making the other type of error.

- To reduce β, ↑ α Value.

- Managers decide the significance level.

- It is more dangerous to accept a false hyp., (Type 2 error than to reject a Correct one) Type 1 error.

- So we keep the prob of committing Type 1 error at a Certain level. Called level of Significance.

- Level of Significance or size of rejection region or size of Critical region or simply Size of test

- At 5% level of Significance, means that the Prob of accepting a true hypothesis 95%.

One-tailed test :

It is called so, 'coz the rejection region will be located in one tail which may be either left or right depending upon the alternative hypothesis formulated.

Ex :-

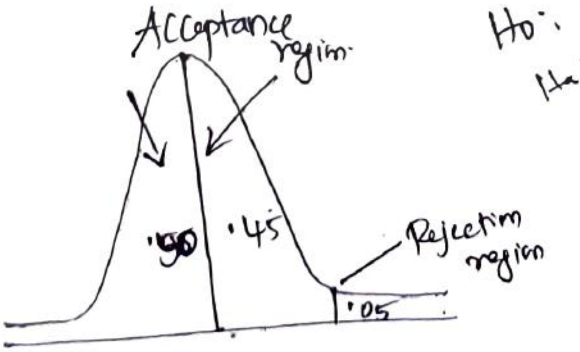
Testing a hypothesis that the average income per household is greater than Rs. 1000, against the alternative hypothesis that the income is Rs. 1000, we will place all the all the alpha risk on right side of the sampling distbn & test will be one-sided right test. $H_0 : \mu > 1000$

$H_a : \mu = 1000$

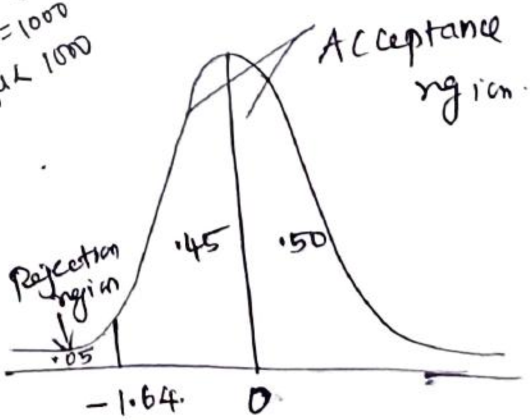
Testing a hypothesis that the average income/household is Rs. 1,000 against alternative that the income is less than Rs. 1000 or less, alpha risk on left side of sampling distbn & the test will be one-sided left tail test.

$H_0 : \mu = 1000$

$H_a : \mu < 1000$



one-tail $\alpha = 0.05$



population mean μ_0

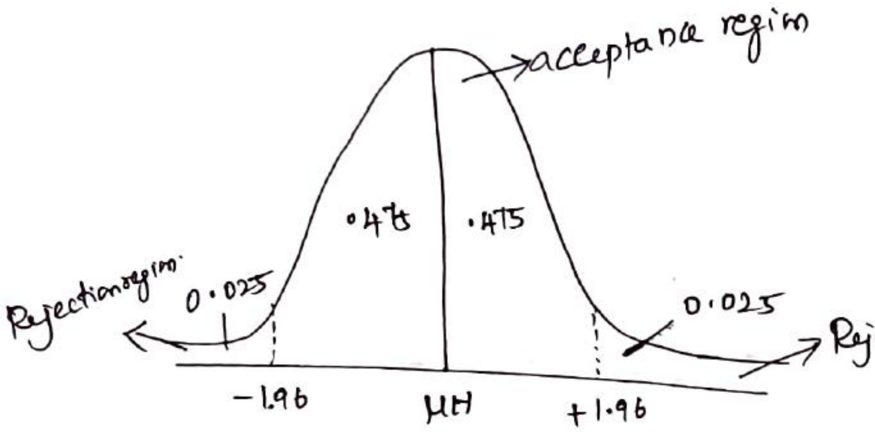
$H_0 : \mu = \mu_0$

$H_a : \mu \neq \mu_0$ ($\mu > \mu_0$ or $\mu < \mu_0$) (2-tailed)

(Right) (Left)

$H_a : \mu > \mu_0$, $H_a : \mu < \mu_0$

Two tailed test :

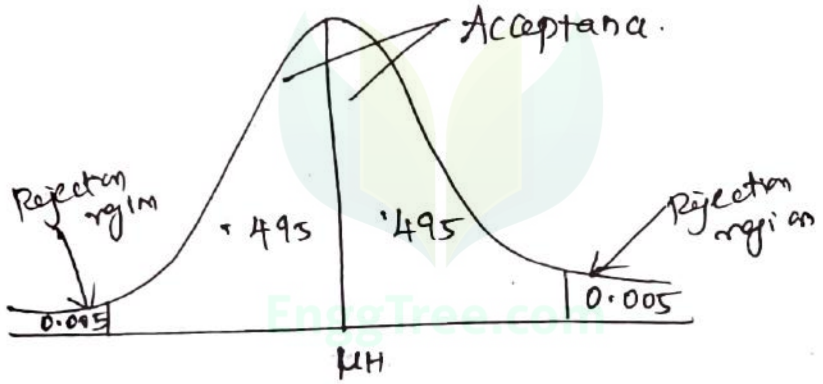


Ho: μ = Mean

H1, μ ≠ Mean

at 0·05

To reduce the risk of committing an error of Type 1. This is done by reducing the size of rejection region.



$\alpha = 0.01$

↓ size of rejection region ↑ prob of accepting our hypothesis

Ex:

The average income / household is Rs. 1000 against the alternative hypothesis that is not Rs.1000 the rejection region would lie on both sides 'coz we would reject the null hypothesis if the mean income in the sample either too far above Rs.1000 or too far below Rs.1000.

20yrs, is being tested with a sample of 40yr old mean.

Soln: $H_0 : \mu \geq 23$ lbs ; $H_1 : \mu < 23$ lbs

Hint:

Increase ; upper tail critical

reduce or decrease : lower tail critical

pblm:

For each situation state whether Ho rejected or retained.

a) Given a one-tailed test, lower tail critical with $\alpha = .01$ and

reject
retain
z

a) $Z = -2.34$

$Z_c = -2.33$

Ho: reject

b) $Z = -5.13$

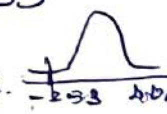
$Z_c = -2.33$

Ho: reject

c) $Z = 4.04$

$Z_c = -2.33$

Ho: ~~rej~~ retain



b) Given a one-tailed test, upper tail critical with $\alpha = .05$ &

accept
reject

d) $Z = 2.00$

$Z_c = 1.~~9~~65$

Ho: reject

e) $Z = -1.80.$

$Z_c = +1.~~9~~65$

Ho: retain

f) $Z = 1.61$

$Z_c = +1.~~9~~65$

Ho: retain

For one tailed tests, the new null hypothesis ·

one-tailed @ lower-tail critical

$H_1: \mu < 500$ then $H_0: \mu \geq 500$ Instead of $H_0: \mu = 500$.

@ upper tail critical.

$H_1: \mu > 500$, then $H_0: \mu \leq 500$.

Ex:

a) An investigator wishes to determine whether, for a sample of drug addicts, the mean score on the depression scale of a personality test differs from a score of 60, is mean score of general population

Soln: $H_0: \mu = 60$; $H_1: \mu \neq 60$

b) To ↑ rainfall, extensive cloud-seeding expts are to be conducted & results are to compared with a baseline fig of 0.54 inch of rainfall.

$H_0: \mu \leq 0.54$; $H_1: \mu > 0.54$.

c) Public health statistics indicate, we will assume, that american males gain an avg of 23 lbs during 20yr period after 40yrs of age. An ambitious weight-reduction pgm spanning

Rejecting Ho is a strong Decision:

Ho is rejected whenever the Zo qualifies as rare outcome - with prob of .05 or less on assumption that Ho is true.

Estimation:

Statistical estimation or estimation is concerned with the methods by which population characteristics are estimated from sample infmn..

-> Finite population - cost of complete censuses may be prohibitive.

-> Infinite population - complete enumerations are impossible.

-> desired degrees of populations

2 types
1) point estimates
2) Interval estimates

Testing a hypothesis about Vitamin C:

- To determine whether Vitamin C increases the intellectual aptitude of high school students. After being randomly selected from some large school district, each of 36 students takes a daily dose of 90 mg of Vit C for a period of 2 months b4 being tested for IQ.

Ordinarily, IQ scores for all students in this school district approximate a nml., distbn with a mean of 100 and a std deviation of 15.

Research Pblm:

Does the daily ingestion of Vit C cause an increase, on average in IQ scores among all students in the school dt?

Hypothesis: upper tail critical

$H_0: \mu \leq 100$; $H_1: \mu > 100$

Decision rule:

Reject H_0 at the 0.05 level of significance if $Z \geq 1.65$

Calculations:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{36}} = 2.5;$$

$$Z_0 = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{100-100}{25} = 0.4$$

Critical Z-Values.

Type of test	Level of Significance (α) .05	.01.
Two-tailed or non-directional (Ho : μ = Some no) (H1 : $\mu \neq$ Some no)	± 1.96	± 2.58
one-tailed or directional test, lower tail critical. Ho : $\mu \geq$ Some no H1 : $\mu <$ Some no	-1.65	-2.33
One-tailed or directional test, upper tail critical Ho : $\mu \leq$ Some no H1 : $\mu >$ some no	$+1.65$	$+2.33$

* Specify the decision rule for each of the foll., situations

a) two-~~test~~tailed test with $\alpha = -0.05$, ± 1.96

Reject Ho @ 0.05 level of Significance if 'z' equals or is more positive than 1.96 or z equals or is more -ve than -1.96.

b) a one tailed test, upper tail critical with $\alpha = .01$

c) a one tailed test, lower tail critical with $\alpha = .05$

d) a two tailed test with $\alpha = .01$

	Accept	Reject
Ho True	CD	T-I
Ho false	T-II	CD

z-test is approriate :

Sample = 36   n>30

4 possible outcomes!.

i) If Ho is really true. ('coz Vit C does not cause an ↑ in the population mean IQ), then it is a correct decision to retain the true Ho. In this case, we could conclude correctly that there is no evidence that Vit C ↑ IQ.

ii) If Ho is really true, it is type-I error to reject the true Ho and conclude that Vitamin C ↑ IQ when it doesn't do. This will allow us to chase after something does not exist.

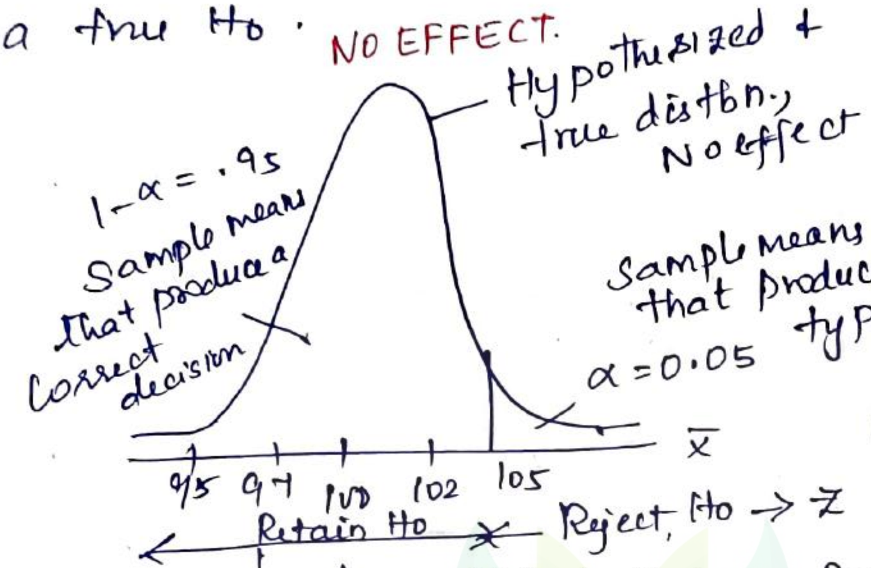
iii) If Ho is really false ('coz Vit C really causes an ↑ in the population mean IQ), then it is a type-II error to retain the false Ho and conclude that there is no evidence that Vit C increases IQ, when infact it does. It is called misses, 'coz they fail to detect a potentially important relationship, such as that b/w Vit C & IQ.

iv) If Ho really is false, then it is a correct decision to reject the false Ho & conclude that Vit C ↑ IQ.

If Ho Really is true.

'Coz Vit c' doesn't ↑ the population mean IQ.

In this case, we concern about retaining or rejecting a true Ho.


NO EFFECT.
Hypothesized + true distbn., No effect
1-α = .95
Sample means that produce a correct decision
Sample means that produce a type 1 error
α = 0.05
$\bar{x}$
95
97
100
102
105
Retain Ho
Reject Ho → Z

('Coz Vit C causes no ↑es in IQ)

- True sampling distribution - from which the one observed sample mean actually originates.

∴, the one observed sample mean in the expt can be viewed as being random selected from the the hypothesized distribution.

- If Ho is really true, the prob of type1 error, α equals the level of significance and the probability of a correct decisions equals 1-α.

- The level of significance indicates the proportion of the total area of the sampling distribution in the rejection region for Ho.

Reducing the prob of T-I error:

- If Ho is really true, the present test will produce a Correct decision with a prob of .95 & Type I error with a prob of .05.

- If a false alarm has serious consequences, the prob of a type I error can be reduced to .01 or even to .001 simply by using .01 or .001 level of significance.

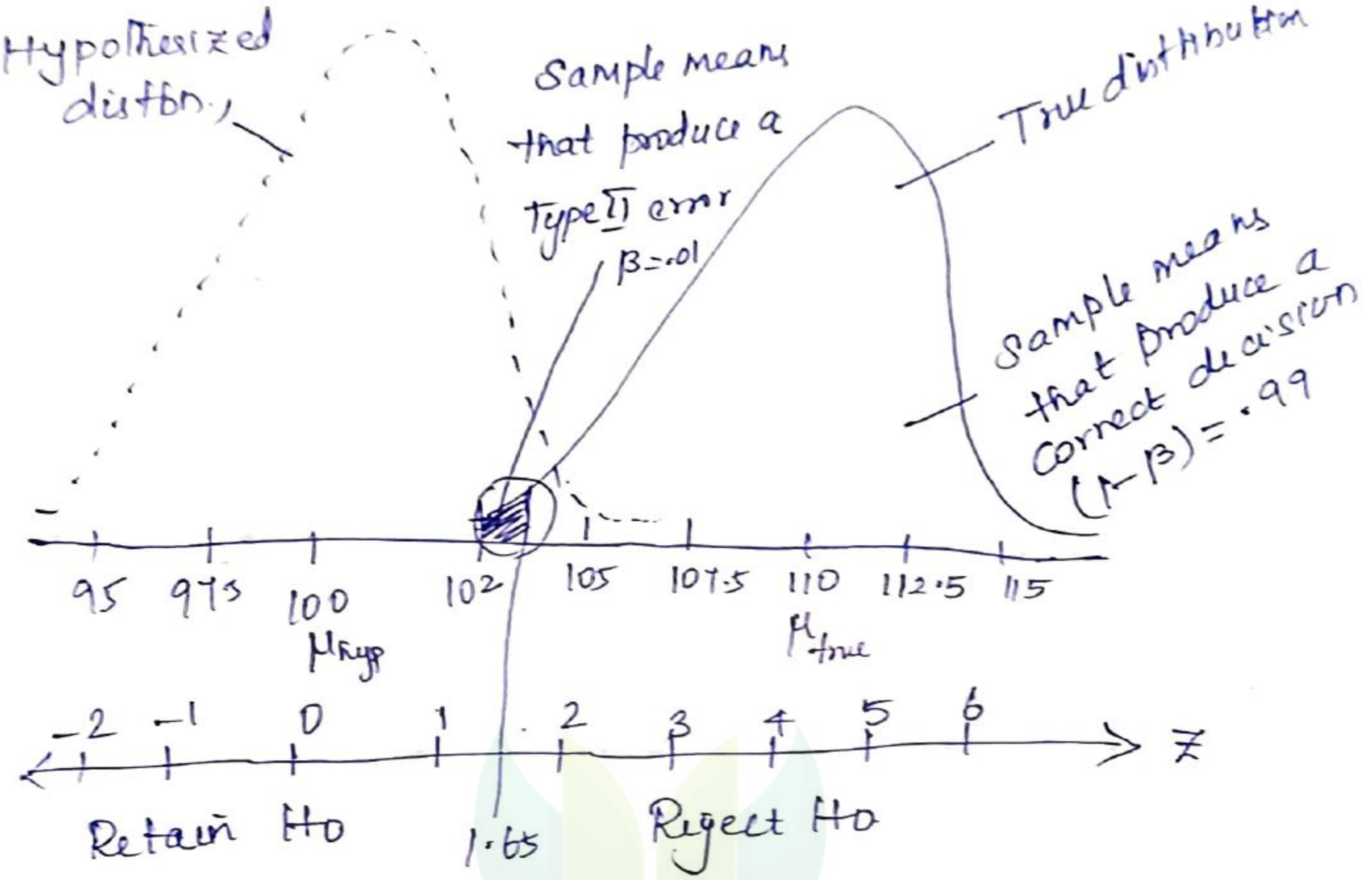
- Any one of the level of significance is preferred, for the Vit C test If, for instance, a false alarm could cause the adoption of an expensive pgm to supply worthless Vit C to all students in the district and, perhaps, the creation of an accelerated curriculum to accomodate the fictitious increase in intellectual aptitude.

If Ho Really is False, because of a large effect:

If Ho is really false / coz Vit C ↑es the population mean by not just few pts, but by many points for ex 10 pts.

→ Any difference b/w true & hypothesized ~~sampling~~ ~~distbn~~ population mean is effect.

Large effect ·



- It is essential to distinguish b/w hypothesized sampling diston and the true sampling diston.
- Centered about the hypo thesized population mean of 100, it serves as a parent diston for the familiar decision rule with a critical z of 1.65 for the projected one-tailed test.

True Sampling distribution:

If mean is ↑ by 10 points then true sampling diston., serves as the parent distribution for the one randomly selected sample mean that will be observed in the expt.

- Viewed relative to the decision rule, the one randomly selected sample mean dictates whether we retain or reject the false Ho.

Low probability of a Type II error for a Large effect:

- A randomly selected sample mean Originates from the very small black portion of the true sampling distbn. of the mean, its 'Z' value is less than 1.65 and ∴ Ho is retained.
- If E Ho is really false, this is an incorrect decision or type-II error - a miss announced as a lack of evidence that Vit C ↑ IQ, even though it does.
- The prob of a type-II error, Symbolized by the letter (β) equals 0.01

High prob of a correct Decision for a large effect.

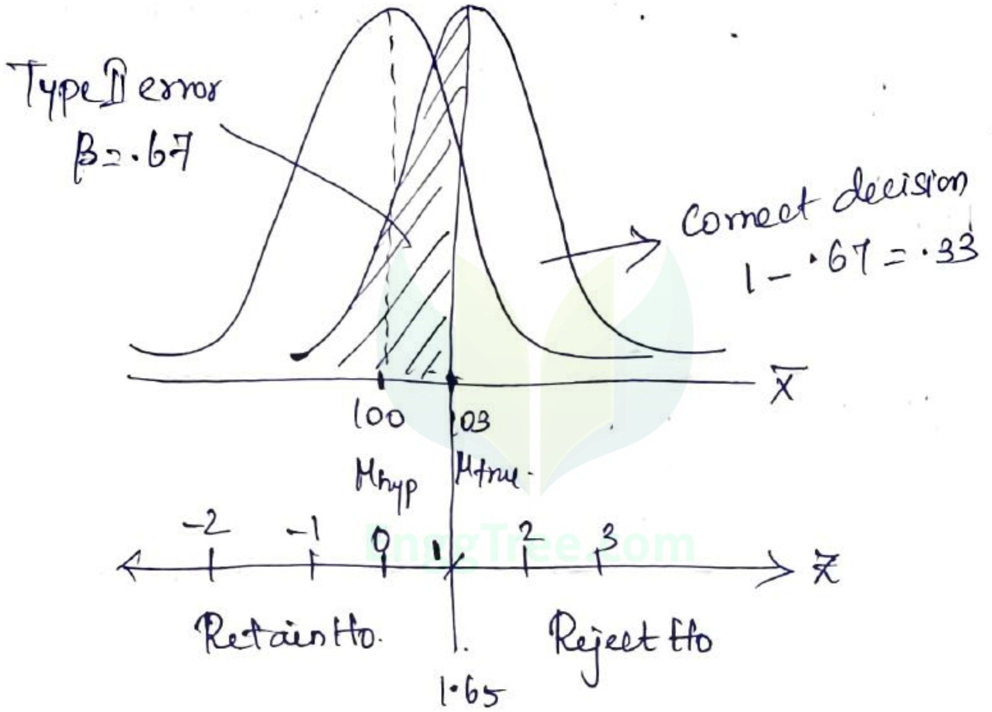
- When a sample mean Originates from the large shaded portion of the true sampling distbn, its Z value equals or exceeds 1.65 and Ho is rejected.
- Ho really is false, then it is a Correct decision announced as evidence that Vit 'C' increases IQ.
- Then the prob of Correct decision 1-β = .99

→ One sided test performs well.

If Ho really is False, because of a small effect.

- Case study for Vit C increases the pop, mean IQ by only a few points - 3 pts.
- Then Hypothesis Sampling Distbn - Centered at 100

  True sampling distbn - Centered at 103.



* Low prob of a correct decision for a small effect.

→ W.r.t the decision rule, the true sampling distbn supplies two types of randomly selected sample means:

i) that produce type II error, coz they Originate from shaded region.

ii) those that produce a correct decision coz they Originate from the white part.

- Because of smaller 5-point effect, the true and hypothesized population means are much closer.
- As a result, the entire true sampling distbn. is shifted towards the retention region for the false Ho.
- Now the projected one-tailed test performs more poorly; there is high probability of 0.67 that a type II error will be committed & low-probability 0.33 that correct decision will be made.

(7.) If Ho is really false, the probability of a type II error, β and the prob of correct decision (1-β) depend on the size of the effect, that is diff, b/w the true and the hypothesized population means. The smaller the effect, the higher the probability of a type II error and the lower the probability of a correct decision.

Influence of Sample size:

→ w.r.t the previous ex. the investigator might not be too concerned about the low detection rate of 0.33 for the relatively small three-point effect of Vit C on IQ.

→ But Special cases make us worry about it.

→ for the previow expt, Vit C has many positive effects including the reduction in the duration and severity of common colds and no side effects. furthermore, huge quantities of Vit C might be available at no cost to the school district.

→ Another point, a fairly mild one such as a small increase in the pop., mean IQ, might clinch the Supply of Vit C. to all the students in the district.

• To increase the prob of detecting a false Ho, ↑ the Sample size.

→ Assume still the expt has 3-point effect on IQ, we can check the properties of the projected one-tailed test when the sample size is increased from 36 to 100 students.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

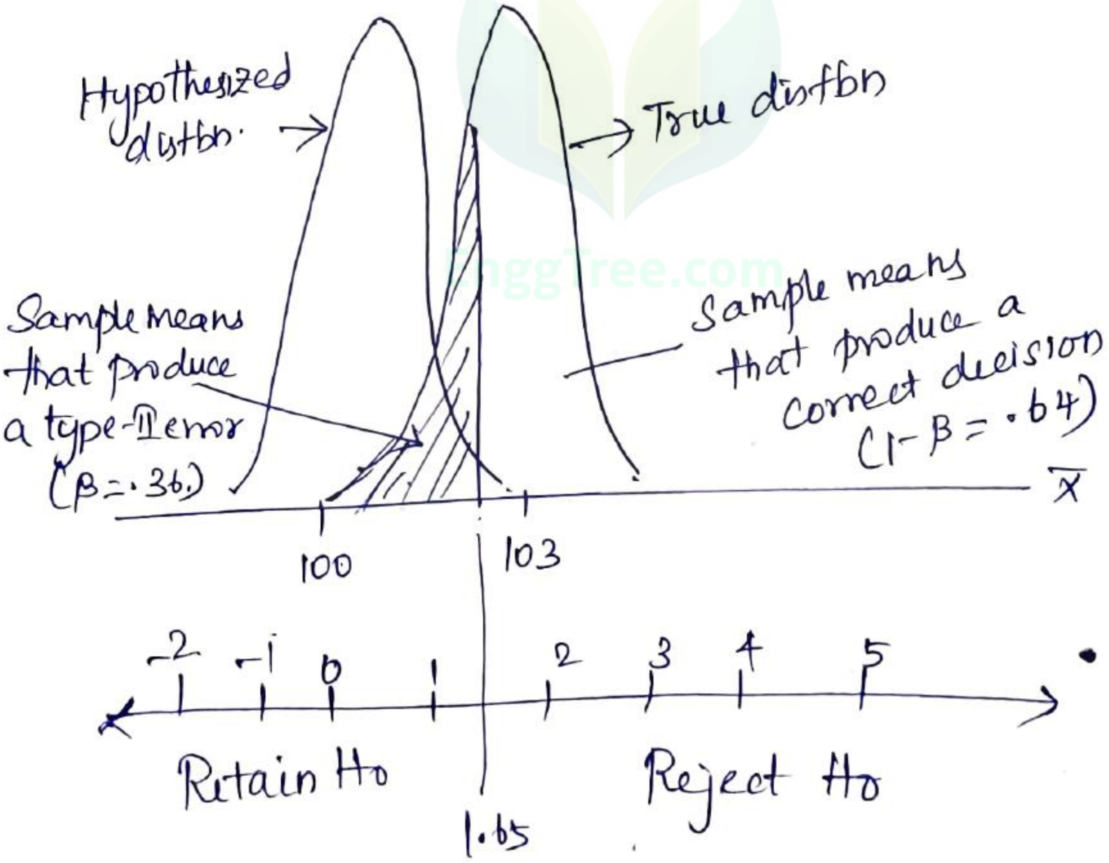
for the original experiment with its sample size of 36,

$$\sigma_{\bar{x}} = \frac{15}{\sqrt{36}} = \frac{15}{6} = 2.5$$

for $n=100$

$$\sigma_{\bar{x}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1.5$$

When 'n' ↑ '$\sigma_{\bar{x}}$' ↓.


Hypothesized distbn.
True distbn
Sample means that produce a type-II error (β=.36)
Sample means that produce a correct decision (1-β=.64)
x̄
100
103
-2
-1
0
1
2
3
4
5
Retain Ho
Reject Ho
1.65

~~Draw back~~

Drawbacks of reducing $\sigma_{\bar{X}}$.

1) It shrinks the upper retention region back toward the hypothesized population mean of 100.

2) It shrinks the entire true sampling distbn toward the true population mean of 103.

From the figure, it is understood for 100 students, fewer sample means (.36) produce a type II error 'coz they originate from the shaded portion & correct decision (.64) lead to detection of Ho.

→ If $\sigma_{\bar{X}}$ is further reduced, the upper retention region shrinks to the immediate vicinity of the hypothesized population mean of 100 and the entire true sampling distbn., of the mean, shrinks to the immediate vicinity of the true population mean of 103. The net result is that type II error hardly ever is committed and the small 3-pt effect virtually is detected.

Power and sample size :

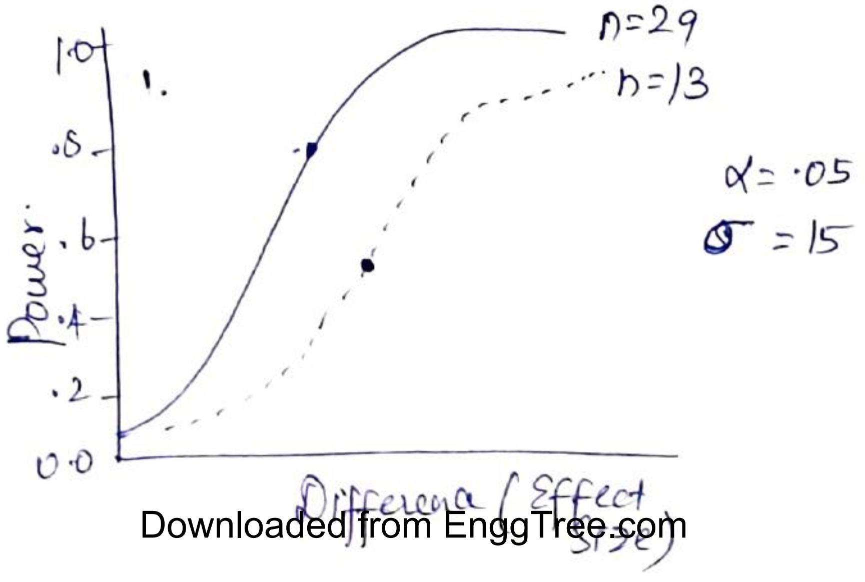
The power of a hypothesis test equals the prob $(1-\beta)$ of detecting a particular effect when the null hypothesis is False.

- power is simply the Complement of $(1-\beta)$ of the prob., 'β' of failing to detect the effect.

- Selection of sample size should reflect
  i) the smallest important effect.
  ii) reasonable degree of power for detecting that effect.

Power curves:

- It shows ~~the~~ how the likelihood of detecting any possible effect - ranging from very small to very large - varies for a fixed sample size.



The dot on the power curve for a sample of 13 indicates that 7-pt effect will be detected with the power of approximately 0.50.

Point Estimate for μ:

A point estimate for μ uses a single value to represent the unknown population mean.

For SAT Score example,

$$\mu = 533$$

↳ point estimate.

→ Drawback is inaccurate.

→ So we use alternative interval estimates or confidence interval.

Confidence interval (CI) for μ.

CI for 'μ' uses a range of values that with a known degree of certainity, includes the unknown population mean.

Why Confidence Intervals Work?

3 important properties

i) The mean of the SD equals the unknown population mean for all local freshman.

ii) The Std, error of the sampling distribution equals the value

iii) The shape of the sampling distbn approximates a normal distribution because the sample size of 100 satisfies the reqts, of central limit theorem.

Series of confidence Interval:

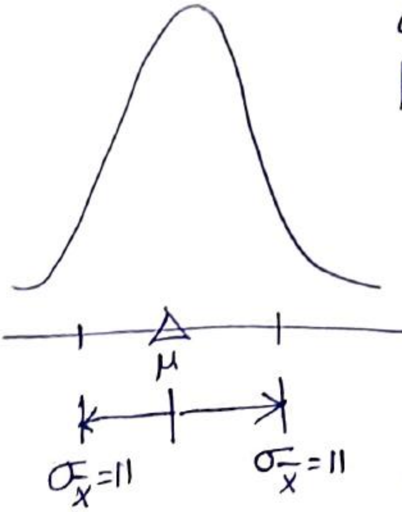
→ Practically, only one sample mean is actually taken from sampling distbn. and used to construct a single 95% CI.

→ Imagine a series of samples from sampling distbn, For each sample mean, construct a 95% CI by adding $1.96\,\sigma_{\bar{x}}$ to the sample mean & subtracting $1.96\,\sigma_{\bar{x}}$ from the sample mean.

$$\bar{X} \pm 1.96\,\sigma_{\bar{X}}$$

to obtain a 95% confidence interval for each sample mean.

True Confidence Interval:

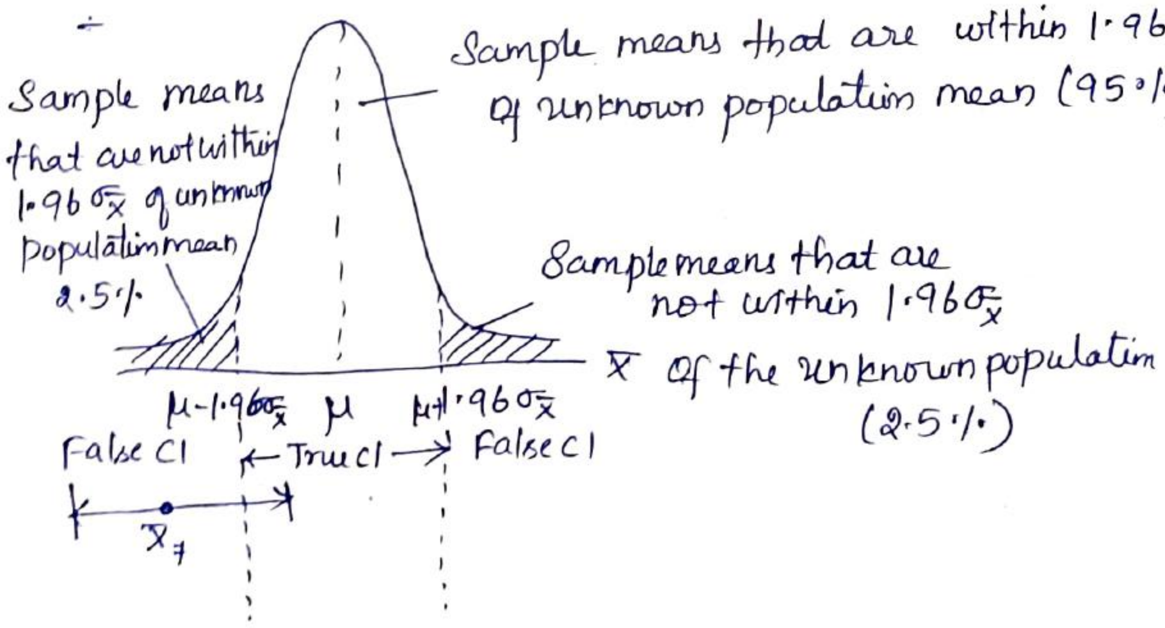


do 95% of these CI include the unknown population mean?

From the figure, Sampling distbn. is normal 95% of all sample means are within 1.96 Std., errors of the unknown population mean (i.e) 95% of all sample means deviate less than 1.96 Standard errors from the unknown population mean.

False Confidence interval:

- 5% of all confidence intervals fail to include the unknown population mean.


Sample means that are within $1.96\sigma_{\bar{x}}$ of unknown population mean (95%)
Sample means that are not within $1.96\sigma_{\bar{x}}$ of unknown population mean 2.5%
Sample means that are not within $1.96\sigma_{\bar{x}}$
$\bar{x}$ of the unknown population (2.5%)
$\mu-1.96\sigma_{\bar{x}}$
μ
$\mu+1.96\sigma_{\bar{x}}$
False CI
True CI
False CI
$\bar{x}_7$

CI for μ based on Z.

$$\bar{x} \pm (Z_{conf})(\sigma_{\bar{x}})$$

Given $\bar{x} = 533$, $Z_{conf} = 1.96$, $\sigma_{\bar{x}} = 11$

$$533 \pm (1.96)(11) = 533 \pm 21.56 = \begin{cases} 554.56 \\ 511.44 \end{cases}$$

When the level of confidence equals 95% or more we can be reasonably confident that the one observed confidence interval include the true population mean

- The level of confidence indicates the percent of time that a series of CI includes the unknown population characteristic such as the population mean.

For 99% $Z_{conf} = \pm 2.58$

$$533 \pm (2.58)(11) = 533 \pm 28.38 = \begin{cases} 561.38 \\ 504.62 \end{cases}$$

Effect on width of interval:

- at 99% CI of 504.62 to 561.38 is wider and ∴ less precise than the corresponding 95% confidence interval of 511.44 to 554.56.
- The shift from a 95% to a 99% level of confidence requires an increase in the value of Z_{conf} from 1.96 to 2.58. This increase causes a wider, less precise confidence interval.

Choosing a level of Confidence:

→ Although many different levels of confidence have been used, 95% and 99% are the most prevalent.

→ a larger level of Confidence, such as 99% Should be reserved for Situations in which a false interval might have particularly serious consequences.

Effect of Sample Size:

→ The larger the sample size, the smaller the $\sigma_{\bar{x}}$ and hence the more precise the Confidence interval will be.

Hypothesis Tests or Confidence intervals?

It indicates whether or not an effect is present, whereas confidence intervals indicate the possible size of the effect.

Confidence interval for population percent:

Sample percent ± (1.96) (Std., error of the percent)

Sample size and Margin of Error:

If 'n' is high, Margin of Error is less.

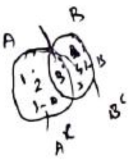
for 1500, ±3 margin of Error.

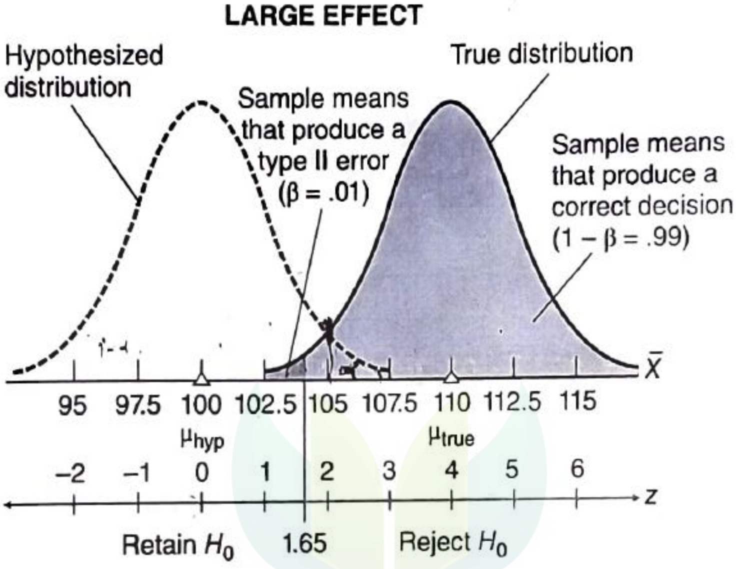
500, ±5 "

100, ±10 "

UNIT II -PROBLEMS

1. Indicate whether the following statements, all referring to Figure 11.4, are true or false: (a) The assumption that H0 really is false is depicted by the separation of the hypothesized and true distributions. (b) In practice, when actually testing a hypothesis, we would not know that the true population mean equals 110. (c) The one observed sample mean is viewed as originating from the hypothesized sampling distribution. (d) A correct decision would be made if the one observed sample mean has a value of 103.





Solution:

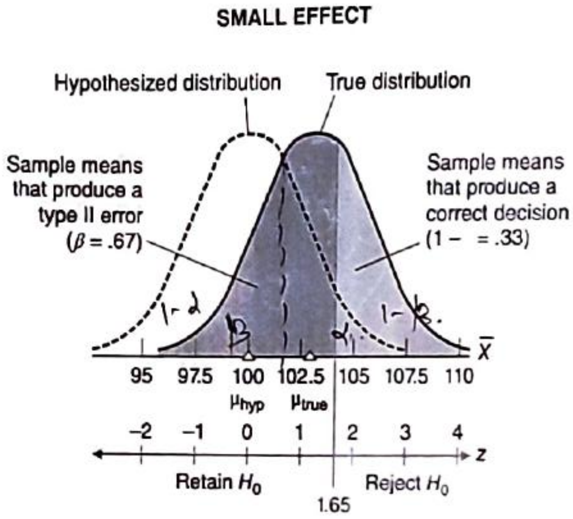
(a) True

(b) True

(c) False. The one observed sample mean originates from the true sampling distribution.

(d) False. If the one observed sample mean has a value of 103, an incorrect decision would be made because the false H0 would be retained

2. Indicate whether the following statements, all referring to Figure 11.5, are true or false: (a) The value of the true population mean (103) dictates the location of the true sampling distribution. (b) The critical value of z (1.65) is based on the true sampling distribution. (c) Since the hypothesized population mean of 100 really is false, it would be impossible to observe a sample mean value less than or equal to 100. (d) A correct decision would be made if the one observed sample mean has a value of 105.


SMALL EFFECT
Hypothesized distribution
True distribution
Sample means that produce a type II error ($\beta = .67$)
Sample means that produce a correct decision (1 − = .33)
1-α
β
α
1-β.
95
97.5
100
102.5
105
107.5
110
$\bar{X}$
μ_{hyp}
μ_{true}
−2
−1
0
1
2
3
4
z
Retain H_0
Reject H_0
1.65

Solution:

a) True

(b) False. The critical value of z (1.65) is based on the hypothesized sampling distribution.

(c) False. Since the true sampling distribution goes below 100, a sample mean less than or equal to 100 is possible, although not highly likely.

(d) True

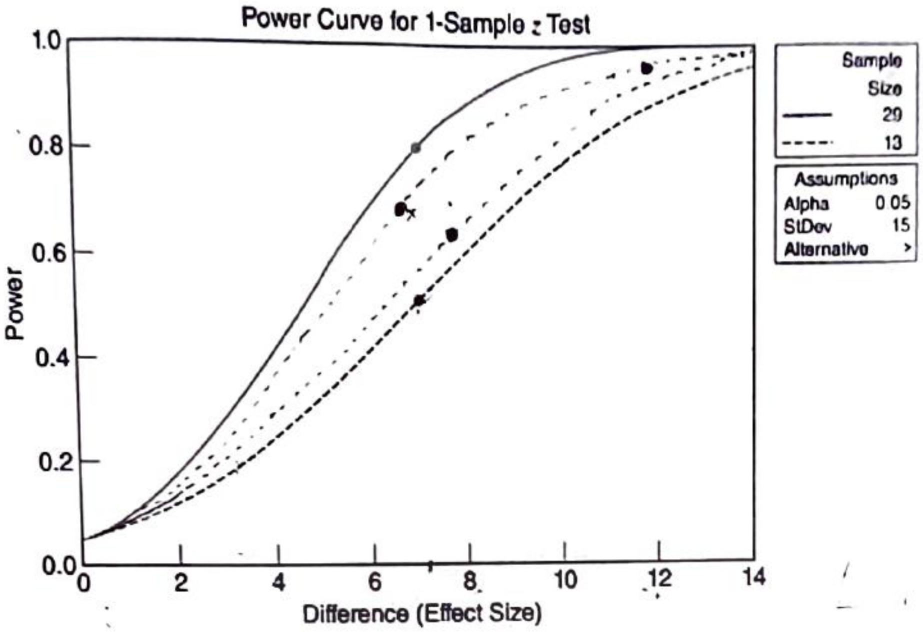
3. Comment critically on the following experimental reports: (a) Using a group of 4 subjects, an investigator announces that H0 was retained at the .05 level of significance. (b) Using a group of 600 subjects, an investigator reports that H0 was rejected at the .05 level of significance.

Solution:

(a) Because of the small sample size, only very large effects will be detected.

(b) Because of the large sample size, even small, unimportant effects will be detected.

4. Consult the power curves in Figure 11.7 to estimate the approximate detection rates, rounded to the nearest tenth, for the following situations: (a) a three-point effect, with a sample size of 29 (b) a six-point effect, with a sample size of 13 (c) a twelve-point effect, with a sample size of 13.



Solution:

(a) .3

(b) .4

(c) .9

5. An investigator consults a chart to determine the sample size required to detect an eight-point effect with a probability of .80. What happens to this detection rate of .80—will it actually be smaller, the same, or larger—if, unknown to the investigator, the true effect actually equals (a) twelve points? (b) five points?

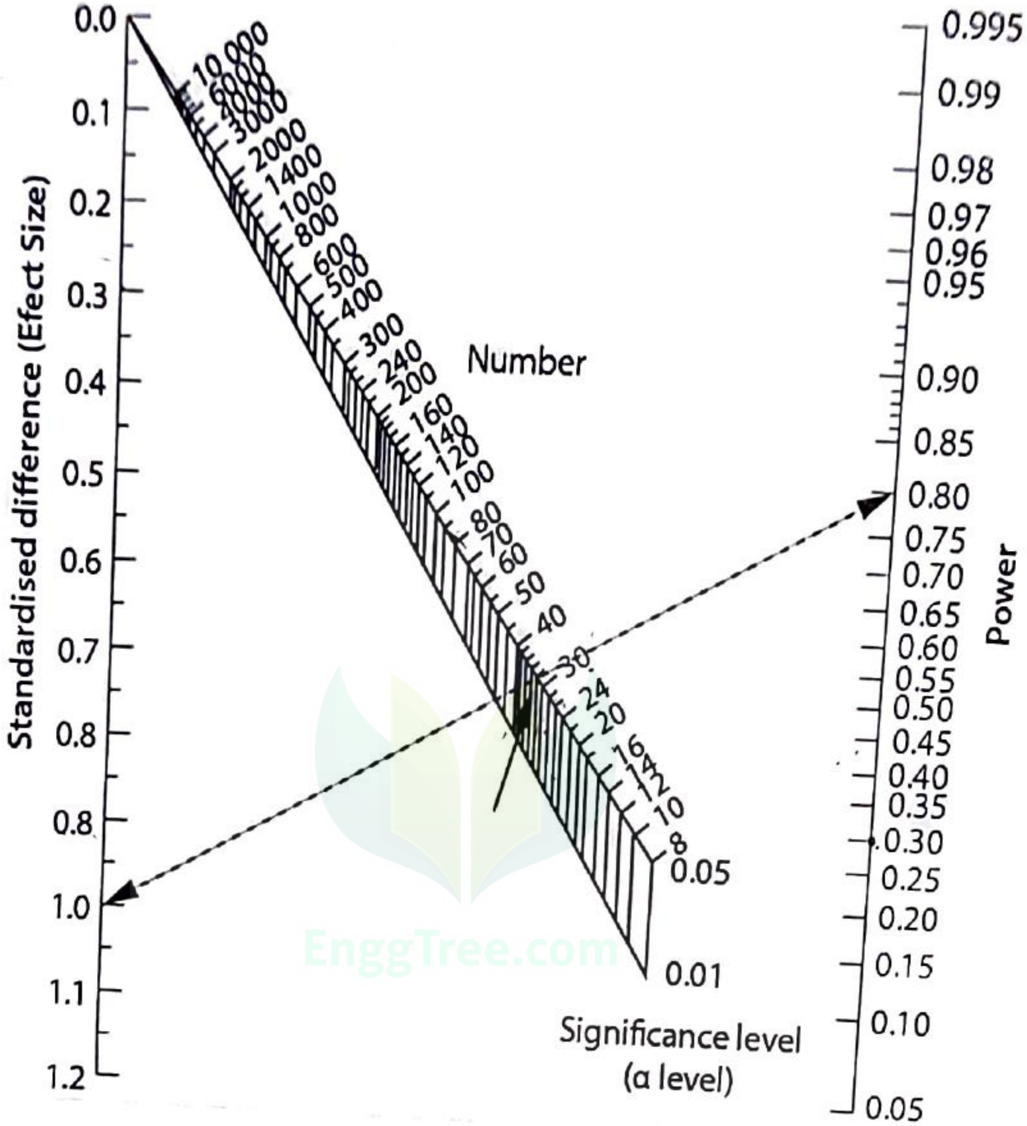
Effect ↑ power ↑.

Solution:

a) The power for the 12-point effect is larger than .80 because the true sampling distribution is shifted further into the rejection region for the false H0.

(b) The power for the 5-point effect is smaller than .80 because the true sampling distribution is shifted further into the retention region for the false H

accept


Standardised difference (Efect Size)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.8
1.0
1.1
1.2
Number
10,000
6000
4000
3000
2000
1400
1000
800
600
500
400
300
240
200
160
140
120
100
80
70
60
50
40
30
24
20
16
14
12
10
8
0.05
0.01
Significance level
(α level)
Power
0.995
0.99
0.98
0.97
0.96
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05

***11.14** Recalling the vitamin C experiment described in this chapter, you could describe the null hypothesis in both symbols and words as follows:

$H_0 : \mu \leq 100$, that is, vitamin C does not increase IQ

Following the format of Table 11.2 and being as specific as possible, you could describe the four possible outcomes of the vitamin C experiment as follows:

	STATUS OF H_0	
DECISION	**TRUE H_0**	**FALSE H_0**
Retain H_0	*Correct Decision:* Conclude that there is no evidence that vitamin C increases IQ when in fact it doesn't.	*Type II Error:* Conclude that there is no evidence that vitamin C increases IQ when in fact it does.
Reject H_0	*Type I Error:* Conclude that vitamin C increases IQ when in fact it doesn't.	*Correct Decision:* Conclude that vitamin C increases IQ when in fact it does.

Using the answer for the vitamin C experiment as a model, specify the null hypothesis and the four possible outcomes for each of the following exercises:

To increase rainfall, with baseline $\underset{H_0}{\mu \leq 0.5}$

11.14 (a) $H_0: \mu \leq 0.54$, that is, cloud seeding has no effect on rainfall.

	STATUS OF H_0	
DECISION	**TRUE H_0**	**FALSE H_0**
Retain H_0	*Correct Decision:* Conclude that there is no evidence that cloud seeding increases rainfall when in fact it does not.	*Type II Error:* Conclude that there is no evidence that cloud seeding increases rainfall when in fact it does.
Reject H_0	*Type I Error:* Conclude that cloud seeding increases rainfall when in fact it does not.	*Correct Decision:* Conclude that cloud seeding increases rainfall when infact it does.

6. A production line at a candy plant is designed to yield 2-pound boxes of assorted candies whose weights in fact follow a normal distribution with a mean of 33 ounces and a standard deviation of .30 ounce. A random sample of 36 boxes from the production of the most recent shift reveals a mean weight of 33.09 ounces. (Incidentally, if you think about it, this is an exception to the usual situation where the investigator hopes to reject the null hypothesis.)

(a) Describe the population being tested.

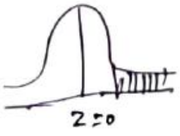
(b) Using the customary procedure, test the null hypothesis at the .05 level of significance.

$H_0 : \mu = 33$
$H_1 : \mu \neq 33$

(c) Someone uses a one-tailed test, upper tail critical, because the sample mean of 33.09 exceeds the hypothesized population mean of 33. Any comment?

$\mu = 33; \; n = 36; \; \bar{x} = 33.09$

Solution: $H_0 : \mu \leq 33, \; H_1 : \mu > 33$



$z = 33.09 -$

7. Consult the power curves in Figure 11.7 to estimate the approximate detection rate, rounded to the nearest tenth, for each of the following situations: (a) a four-point effect, with a sample size of 13 (b) a ten-point effect, with a sample size of 29 (c) a seven-point effect with a sample size of 18 (Interpolate).

Solution:

8. A random sample of 200 graduates of U.S. colleges reveals a mean annual income of $62,600. What is the best estimate of the unknown mean annual income for all graduates of U.S. colleges?

Solution:

$62,600

9. Reading achievement scores are obtained for a group of fourth graders. A score of 4.0 indicates a level of achievement appropriate for fourth grade, a score below 4.0 indicates underachievement, and a score above 4.0 indicates overachievement. Assume that the population standard deviation equals 0.4. A random sample of 64 fourth graders reveals a mean achievement score of 3.82. (a) Construct a 95 percent confidence interval for the unknown population mean. (Remember to convert the standard deviation to a standard error.) (b) Interpret this confidence interval; that is, do you find any consistent evidence either of overachievement or of underachievement?

Solution:

(a) $3.82 \pm 1.96 .4\ 64 = 3.92\ 3.72$

(b) We can claim, with 95 percent confidence, that the interval between 3.72 and 3.92 includes the true population mean reading score for the fourth graders. All of these values suggest that, on average, the fourth graders are underachieving.

10. Before taking the GRE, a random sample of college seniors received special training on how to take the test. After analyzing their scores on the GRE, the investigator reported a dramatic gain, relative to the national average of 500, as indicated by a 95 percent confidence interval of 507 to 527. Are the following interpretations true or false? (a) About 95 percent of all subjects scored between 507 and 527. (b) The interval from 507 to 527 refers to possible values of the population mean for all students who undergo special training. (c) The true population mean definitely is between 507 and 527. (d) This particular interval describes the population mean about 95 percent of the time. (e) In practice, we never really know whether the interval from 507 to 527 is true or false. (f) We can be reasonably confident that the population mean is between 507 and 527.

Solution:

(a) False. We can be 95 percent confident that the mean for all subjects will be between 507 and 527.

(b) True

(c) False. We can be reasonably confident—but not absolutely confident—that the true population mean lies between 507 and 527.

(d) False. This particular interval either describes the one true population mean or fails to describe the one true population mean.

(e) True

(f) True

11. On the basis of a random sample of 120 adults, a pollster reports, with 95 percent confidence, that between 58 and 72 percent of all Americans believe in life after death. (a) If this interval is too wide, what, if anything, can be done with the existing data to obtain a narrower confidence interval? (b) What can be done to obtain a narrower 95 percent confidence interval if another similar investigation is being planned?

Solution:

(a) Switch to an interval having a lesser degree of confidence, such as 90 percent or 75 percent.

(b) Increase the sample size.

12. In a recent scientific sample of about 900 adult Americans, 70 percent favor stricter gun control of assault weapons, with a margin of error of ±4 percent for a 95 percent confidence interval. Therefore, the 95 percent confidence interval equals 66 to 74 percent. Indicate whether the following interpretations are true or false: (a) The interval from 66 to 74 percent refers to possible values of the sample percent. (b) The true population percent is between 66 and 74 percent. (c) In the long run, a series of intervals similar to this one would fail to include the population percent about 5 percent of the time. (d) We can be reasonably confident that the population percent is between 66 and 74 percent.

Solution:

(a) False. The interval from 66 to 74 percent refers to possible values of the population proportion.

(b) False. We can be reasonably confident—but not absolutely confident—that the true population proportion is between 66 and 74 percent.

(c) True

(d) True

13. In Question 10.5 on page 191, it was concluded that, the mean salary among the population of female members of the American Psychological Association is less than that ($82,500) for all comparable members who have a doctorate and teach full time. (a) Given a population standard deviation of $6,000 and a sample mean salary of $80,100 for a random sample of 100 female members, construct a 99 percent confidence interval for the mean salary for all female members. (b) Given this confidence interval, is there any consistent evidence that the mean salary for all female members falls below $82,500, the mean salary for all members?

Solution:

(a) 80,100 ± 2.58 6,000 100 = 81,648 78,552

(b) We can claim, with 99 percent confidence, that the interval between $78,552 and $81,648 includes the true population mean salary for all female members of the American Psychological Association. All of these values suggest that, on average, females' salaries are less than males' salaries.

14. Imagine that one of the following 95 percent confidence intervals estimates the effect of vitamin C on IQ scores:

95% CONFIDENCE INTERVAL	LOWER LIMIT	UPPER LIMIT
1	100	102
2	95	99
3	102	106
4	90	111
5	91	98

(a) Which one most strongly supports the conclusion that vitamin C increases IQ scores?
(b) Which one implies the largest sample size?
(c) Which one most strongly supports the conclusion that vitamin C decreases IQ scores?
(d) Which one would most likely stimulate the investigator to conduct an additional experiment using larger sample sizes?

Solution:

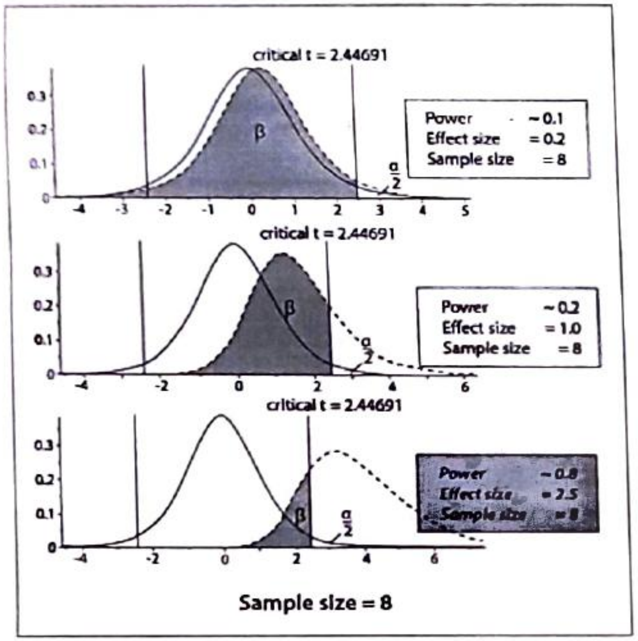
(a) 3

(b) 1

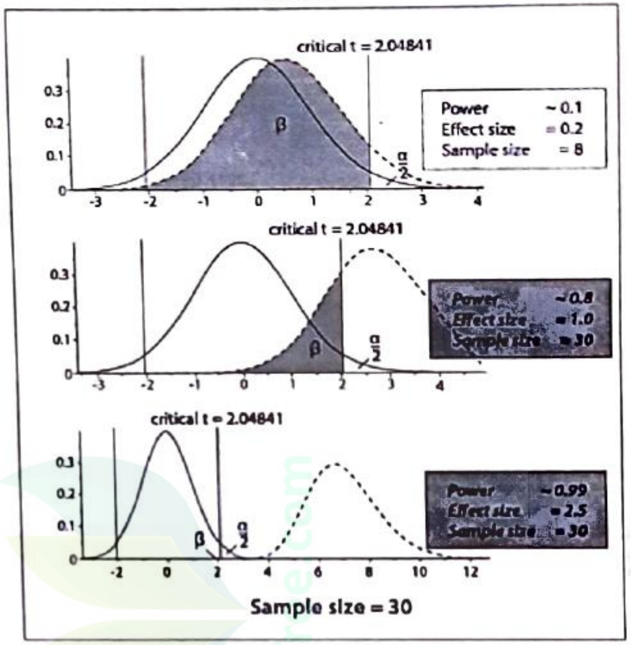
(c) 5

(d) 4

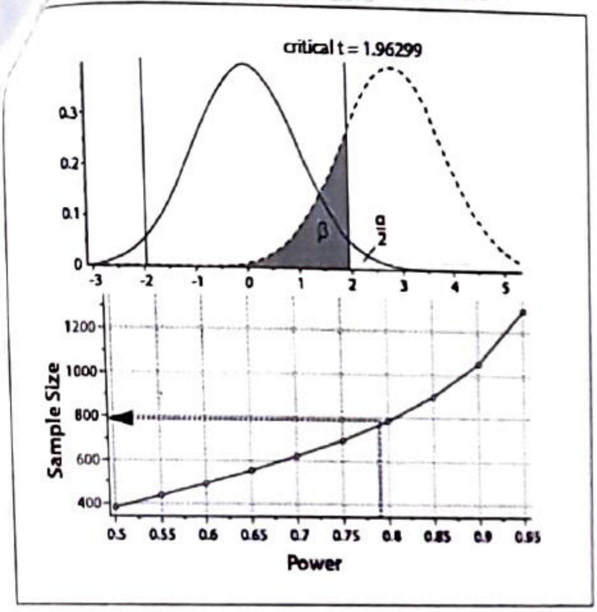
Case 4


critical t = 2.44691
0.3
0.2
0.1
0
β
α/2
-4 -3 -2 -1 0 1 2 3 4 5
Power ~0.1
Effect size = 0.2
Sample size = 8
critical t = 2.44691
-4 -2 0 2 4 6
Power ~0.2
Effect size = 1.0
Sample size = 8
critical t = 2.44691
-4 -2 0 2 4 6
Power ~0.8
Effect size = 2.5
Sample size = 8
Sample size = 8

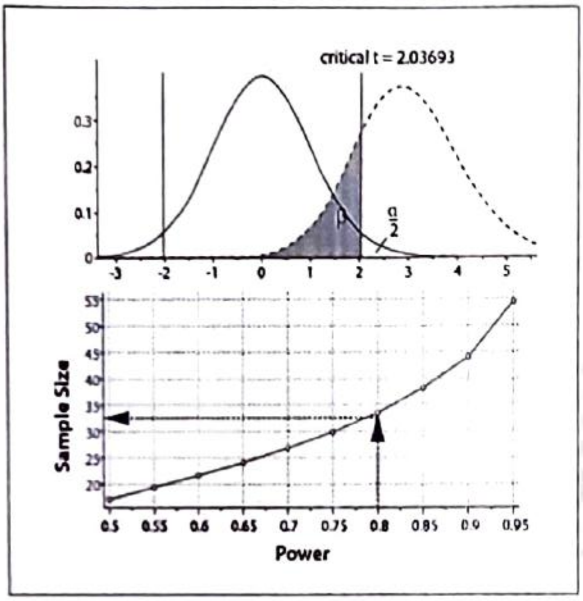
Case 5


critical t = 2.04841
0.3
0.2
0.1
0
β
α/2
-3 -2 -1 0 1 2 3 4
Power ~0.1
Effect size = 0.2
Sample size = 8
critical t = 2.04841
-3 -2 -1 0 1 2 3 4
Power ~0.8
Effect size = 1.0
Sample size = 30
critical t = 2.04841
-2 0 2 4 6 8 10 12
Power ~0.99
Effect size = 2.5
Sample size = 30
Sample size = 30

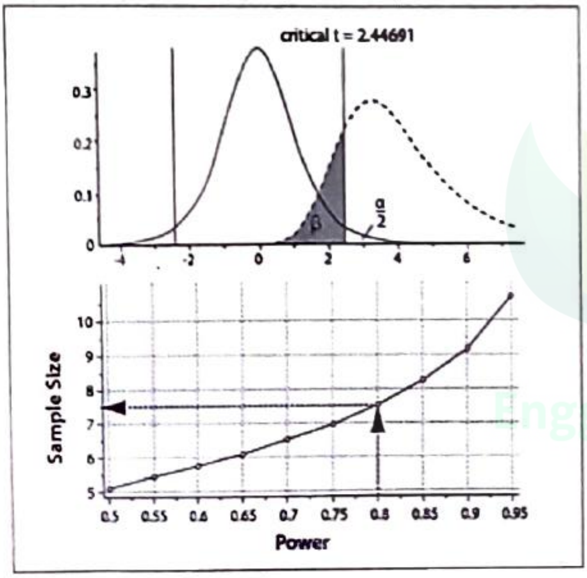
Case 1: *P = 0.8*, ES = 0.2, SS = 788



Case 2: *P = 0.8*, ES = 1, SS = 34



Case 3: *P = 0.8*, ES = 2.5, SS = 8



Unit-III

T-Test

t-test for one sample:

Gas mileage investigation:

- Federal law might eventually specify that new automobiles must average, for ex 45 miles per gallon of gasoline.
- But its complicated to test all car.
- So we take random samples from the entire ~~popu~~ production of each car model.
- If a hypothesis test indicates substandard performance, the manufacturer would be penalized, we will assume $200/car for the entire production.

$$H_0 : \mu \geq 45$$
$$H_1 : \mu < 45 .$$

- From manufacturer's perspective a, type I error is very serious. So we use less 'α' value say 0.01 instead of a customary ·05 level of significance.
- From Federal regulators perspective, a type II error is also serious.

So to reduce Type II error sample size selected to control type-II error that is to ensure a reasonable detection rate for the smallest decline of the truepop, mean below the mandated 45mpg.

- To simplify the Computations in the present ex, the projected one-tailed test is based on data from a very samll sample of only 6 randomly selected cars.
- So we need to replace the Z-test with some t-test that can work will with small sample size.

Sampling Distribution of t:

It represents the distbn, that would be obtained if a value of 't' were calculated for each sample mean for all possible random samples of a given size from some population.

- William Gosset discovered "Student's distbn".
- Each t distribution is associated with a special no called "Degrees of freedom".
- The concept ~~of~~ Degrees of freedom is introduced 'coz we are using variability of sample to estimate the unknown variability in the population.

Degrees of Freedom

$$df = n-1.$$

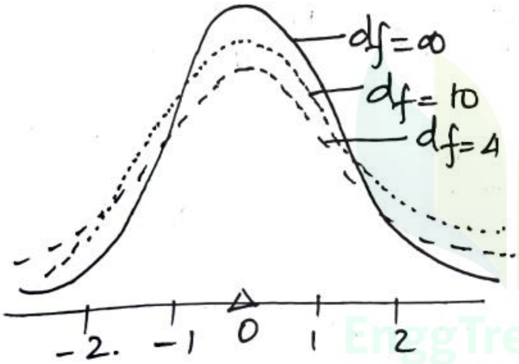
where df is Degrees of Freedom.

n is Sample size.

for gas mileage investigation.

$$df = 6-1=5$$

Compared to standard Normal Distribution.



@ $df=\infty$, the $S_{\bar{x}} = \sigma$, then 't' distbn = Std. nml distbn 'z'.

- Symmetrical, unimodal & bell-shaped with a dense concentration that peaks in the middle and tapers off both right and left of the middle..

- The inflated tails of the 't' distribution particularly apparent with small values of 'df', that constitute the most important difference b/w 't' & 'z' distribution.

P-1. Find the critical 't' values for the foll., tests

a) two-tailed test $\alpha = \cdot 05$, $df=12$

-2.539

b) one-tailed test, @ lower tail critical $\alpha = \cdot 01$, $df=19$

c) " " " upper " " $\alpha = \cdot 05$, $df=38$

d) two tailed test $\alpha = 0.01$, $df=48$.

t- Test:

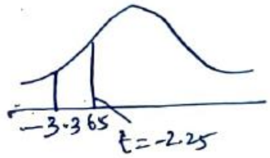
Here population std., deviation is unknown and must be estimated from the sample.

$$t = \frac{\text{Sample mean} - \mu_{hyp}}{\text{estimated std., error}} = \frac{\overline{X} - \mu_{hyp}}{S_{\overline{X}}}$$

with its 't' sampling distbn & $(n-1)$ degrees of Freedom.

for gas mileage pblm, $\overline{X} = 43$, $\mu_{hyp} = 45$; $S_{\overline{X}} = 0.89$.

$$t = \frac{43-45}{0.89} = -2.25$$



with df=5. Observed value of 't' is less negative than the $t_c = -3.365$, so Ho retained and conclude that auto manufacturer Should'nt penalized ∵ the mean gas mileage for the pop., car would be equal the mandated 45 mpg.

Greater variability of t Ratio:

t_c @ 0.01 ⇒ -3.365 (lower - critical)

z_c @ 0.01 ⇒ -2.33 (lower - critical)

Common theme of Hypothesis tests:

→ all of these hypothesis tests represent variations on the same theme: If some observed Characteristic Such as the mean for a random sample, qualifies as a rare outcome under the null hypothesis, the hypothesis will be rejected. otherwise, the hypothesis will be retained.

→ Gas mileage data consists of 6 values: 40, 44, 46, 41, 43 and 44.

→ only '5' observation can be varied about their mean of 43 and ∴ provide valid information for purposes of estimation.

→ the concept of degrees of freedom is introduced only 'coz we are using observations in a sample to estimate some unknown Characteristics of the population.

Estimating the standard error ($s_{\bar{x}}$).

If the population std deviation is unknown it must be estimated from the sample.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad ; \quad S_{\bar{x}} = \frac{s}{\sqrt{n}}.$$

$S_{\bar{x}}$ → estimated std, error of mean

n → equals the sample size.

$$S = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{SS}{df}}$$

'S' - sample standard deviation.

df - degrees of freedom.

$$SS = \Sigma(x - \bar{x})^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n}.$$

This is 'estimated std, error of the mean' is used whenever the unknown population standard deviation must be estimated.

Gras mileage investigation.

3 panels
- Panel I
- Panel II
- Panel III

Panel - I - Calculate 'S'.

↳ Calculating sample mean $\overline{X}$

↳ Sample Std., deviation S.

$$SS = \sum x^2 - \frac{(\Sigma x)^2}{n}$$

$$S = \sqrt{\frac{SS}{n-1}}$$

Panel II : Find $S_{\overline{X}}$

↳ Dividing the sample std deviation, 'S' by sq., root of Sample size, n gives the Value for the estimated std., error $S_{\overline{X}}$.

$$S_{\overline{X}} = \frac{S}{\sqrt{n}}$$

Panel-III: Find calculation for t-ratio.

$$t = \frac{\overline{X} - \mu_{hyp}}{S_{\overline{X}}}$$

Confidence interrals for μ based on 't'.

$$\bar{X} \pm (t_{conf})(S_{\bar{X}})$$

Given this value $t_c = 2.571$ @ 99% of CI, $\bar{X} = 43$,

$S_{\bar{X}} = 0.89$.

$$43 \pm (2.571)(0.89) = 43 \pm 2.29 = \begin{cases} 45.29 \\ 40.71 \end{cases}$$

Interpretation:

We can reasonablly confident that the true mean for the entire population of Cars is neither less than 40.71 mpg nor more than 45.29 mpg.

Calculation for the t-test
(Gas mileage investigation)

1) Find $\bar{X}$ and s

$X = 40, 44, 46, 41, 43, 44$

X	X^2
40	1600
44	1936
46	2116
41	1681
43	1849
44	1936

1) $n = 6$

2) $\Sigma X = 258$

3) $\bar{X} = \frac{\Sigma X}{n} = \frac{258}{6} = 43$

4) $\Sigma x^2 = 11118$

6) $$SS = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

$$= 11118 - \frac{(258)^2}{6} = 11118 - \frac{66564}{6}$$

$$= 11118 - 11094 = 24$$

7) $$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{24}{6-1}} = \sqrt{4.8} = 2.19$$

II. Find $s_{\bar{x}}$

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{2.19}{\sqrt{6}} = \frac{2.19}{2.45} = 0.89$$

III. Calculations for t ratio

$$t = \frac{\bar{X} - \mu_{hyp}}{s_{\bar{x}}} = \frac{43-45}{0.89} = \frac{-2}{0.89} = -2.25$$

t-test for two independent samples :

EPO experiment :-

→ Tour de France is a best known bicycle race has seen some cyclists expelled for attempting to enhance their performance by using a Variety of banned Substances, including a Synthetic "blood-doping" hormone, erythropoietin (EPO) that Stimulates the production of oxygen-bearing RBC.

- A mental health investigator at a large clinic wants to determine whether EPO - viewed as a serious therapeutic tool might increase the endurance of severely depressed patients.
- Volunteer patients are randomly assigned to one of the two groups: a treatment group(X_1) that receives a prescribed amount of EPO and a ctrl. group(X_2) that receives a harmless neutral Substance.
- Subsequent endurance Scores are based on the total time, in mins that each patient remains on a rapidly moving treadmill.
- The Statistical analysis focuses on the difference b/w mean endurance scores for the treatment + ctrl. groups.

For Computational convenience, the results for the Current expt, based on 6 endurance scores / Group. Also for computational convenience, endurance scores have been rounded to the nearest minute even though, in practice, they surely would reflect measurement that is more precise.

→ overlap in the Scores for the two groups. The treatment Scores tend to be slightly larger than the ctrl, scores and this tendency is supported by the mean difference of 5 mins in favor of treatment group.

- Qns:

i) How to interpret this tendency?

ii) Is it real & ∴ likely to reappear in a repeat expt as a difference favouring the treatment group?

iii) Is the tendency transitory? or favouring for ctrl, group?

→ Ans:

A t test for two independent samples which evaluates the mean diff of 5 mins relative to its Variability to answer this question.

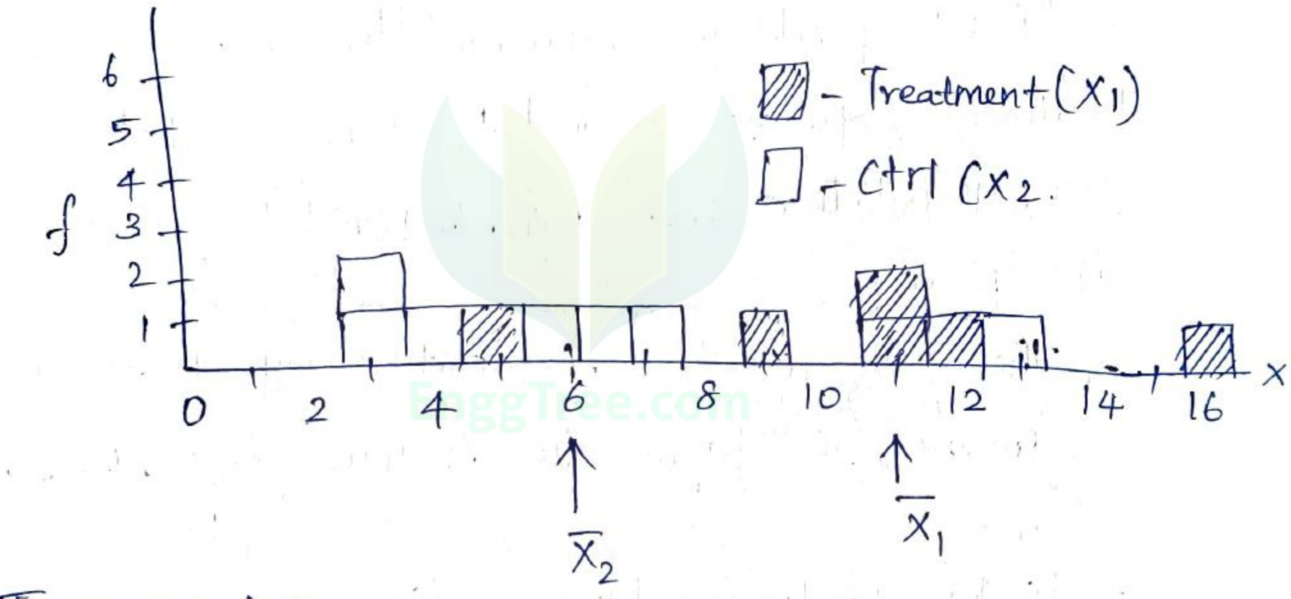
Difference b/w population Means:

- The difference b/w population means reflects the effect of EPo on endurance.
- If EPo has little or no effect on endurance, then the endurance scores would tend to be about the same for both populations of patients and the diff. b/w population means would be close to zero.
- If EPo facilitates endurance, the scores for the treatment population would tend to exceed those for the Ctrl. population and diff b/w population means would be positive.
- The stronger the facilitative effect of EPo on endurance, the larger the positive diff b/w population means.
- If EPo hinders endurance, the endurance scores for the treatment population would tend to be exceeded by those for the Ctrl. population and the difference b/w population means would be -ve.

Two independent samples

→ 2 groups < one treatment, one ctrl., } both different patients.

– When samples are independent, observations in one sample are not paired, on a one-to-one basis with observations in the other sample.



Two population:

– 2 hypothetical population.

– A treatment population:

It is defined for the endurance scores of patients who receive EPO

→ Ctrl, population:

It is defined for the endurance scores of patients who dont receive EPO.

Alter psbl, H_a:

a) directional hypothesis

$$H_1 : \mu_1 - \mu_2 < 0$$

b) non-directional hypothesis

$$H_1 : \mu_1 - \mu_2 \neq 0$$

Sampling Distribution of $\bar{x}_1 - \bar{x}_2$

- It represent the entire spectrum of differences b/w Sample means based on all possible pairs of random samples from the two underlying populations.
- sampling distbn., Centered with μH_0, we find the observed sample mean to be a rare outcome.

Mean of the Sampling distribution: $\mu_{\bar{x}_1 - \bar{x}_2}$

The mean of the sampling distbn of $\bar{x}$ equals the population mean

$$\mu_{\bar{x}} = \mu$$

$\mu_{\bar{x}}$ is the mean of Sampling distbn and μ is the population mean.

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2.$$

LHS is mean of new Sampling distbn,
RHS is diff b/w pop, means

Statistical Hypotheses:

Null Hypothesis:

According to the null hypothesis, nothing special is happening 'coz EPO does not facilitate endurance. In other words, either there is no diff b/w the means for the two populations or the diff, b/w population means is negative

$$H_0: \mu_1 - \mu_2 \leq 0$$

H_0 = null hypothesis

μ_1 & μ_2 are the population means of treatment & ctrl, groups respectively.

Alternative (or) Research Hypothesis:

The investigator wants to reject the null hypothesis only if the treatment increases endurance scores. Given this perspective, the alternative hypothesis sh'ld specify that the diff b/w population means is +ve 'coz EPO facilitates endurance. An equivalent statement is

$$H_1: \mu_1 - \mu_2 > 0.$$

This directionalo hypothesis

Standard error of Sampling Distribution: $\sigma_{\bar{x}_1 - \bar{x}_2}$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

σ_1^2, σ_2^2 are two pop., variances and n_1, n_2 are two sample sizes.

$\sigma_{\bar{x}_1 - \bar{x}_2}$ - new std, error.

t-ratio:

$$t = \frac{\text{diff b/w Sample means} - \text{hypothesized diff b/w Population mean}}{\text{estimated std, error}}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_{hyp}}{S_{\bar{x}_1 - \bar{x}_2}}$$

Panel - I : Calculate SS_1 & SS_2.

$$SS_1 = \Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n_1}$$

$$SS_2 = \Sigma x_2^2 - \frac{(\Sigma x_2)^2}{n_2}$$

Panel-II : find pooled Variance S_p^2

$$S_p^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2} \qquad df = n_1 + n_2 - 2$$

Panel-III :: find std error $S_{\bar{x}_1 - \bar{x}_2}$

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

Panel-IV :: t ratio

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_{hyp}}{S_{\bar{x}_1 - \bar{x}_2}}$$

EPO Experiment:

Research Problem:

Does the population mean endurance score for treatment (EPO) patients exceed that for ctrl, patients?

Statistical hypotheses:

$H_0 : \mu_1 - \mu_2 \geqslant 0 \Rightarrow \mu_1 \geqslant \mu_2$ (There is no significant diff b/w treatment group & ctrl, group)

$H_1 : \mu_1 - \mu_2 < 0 \Rightarrow \mu_1 < \mu_2$. (There is significant diff b/w treatment grp., + ctrl grp)
(Strictly less than 0)

Reject Ho at the 0.05 level of significance if

$t_c = 1.812$; $df = n_1 + n_2 - 2 = 6 + 6 - 2 = 10$.

Calculations:

$$t_0 = \frac{(11-6) - 0}{2.32} = 2.16$$

Decision:

Reject Ho at the 0.05 level of significance bcoz $t_0 = 2.16 > 1.812$.

Interpretation:

The diff., b/w population means is greater than zero. There is evidence that EPO increases that mean endurance scores of treatment patients.

Calculations:

Endurance Scores (mins.)

EPO		Ctrl.	
X_1	X_1^2	X_2	X_2^2
12	144	7	49
5	25	3	9
11	121	4	16
11	121	6	36
9	81	3	9
18	324	13	169
$\Sigma X_1 = 66$	$\Sigma X_1^2 = 816$	$\Sigma X_2 = 36$	$\Sigma X_2^2 = 288$

Panel-I

1) $n_1 = 6$ 2) $n_2 = 6.$ 3) $\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{66}{6} = 11$

4) $\bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{36}{6} = 6$

5) $SS_1 = \Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n_1}$

$= 816 - \frac{(66)^2}{6}$

$= 816 - 726 = 90$

$$SS_2 = \Sigma x_2^2 - \frac{(\Sigma x_2)^2}{n_2}.$$

$$= 288 - \frac{(36)^2}{6}$$

$$= 288 - 216 = 72.$$

Panel-II : Finding the pooled variance S_p^2

$$S_p^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2.} = \frac{90+72}{6+6-2} = \frac{162}{10} = 16.2$$

Panel-III Find $S_{\bar{x}_1 - \bar{x}_2}$

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} = \sqrt{\frac{16.2}{6} + \frac{16.2}{6}} = \sqrt{\frac{32.4}{6}}$$

$$= \sqrt{5.4} = 2.32.$$

Panel-IV finding t_0

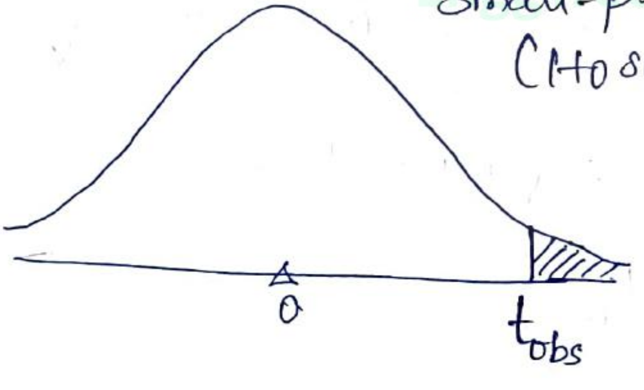
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_{hyp}}{S_{\bar{x}_1 - \bar{x}_2}} = \frac{(11-6)-0}{2.32.}$$

$$= \frac{5}{2.32} = 2.16.$$

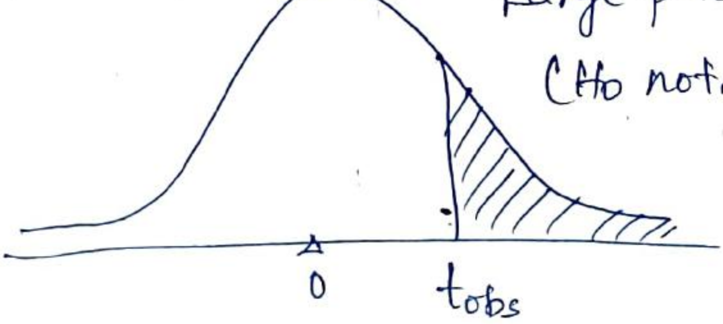
p- values:-

The null hypothesis is neither retained nor rejected but viewed with degrees of suspicion, depending on the degree of rarity of the observed value of t.

The P-value for a test result represents the degree of rarity of that result, given that the null hypothesis is true. Smaller p-values tend to discredit the null hypothesis and to support the Research hypothesis.



Small-p-value
(Ho suspect)



Large p-value
(Ho not suspect)

Finding Approximate p-values :

$P < .05$ and $P > 0.05$

The p-value or prob, value is a number describing how likely it is that your data would have occured by random chance. The level of statistical significance is expressed as p-value b/w 0 and 1.

The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

$P < 0.05$ is statistically significant. It indicates the strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct. ∴, we reject the H_0 and accept the alternative hypothesis.

$P > 0.05$ is not statistically significant and indicates strong evidence for the null hypothesis. This means we retain the null hypothesis and reject the alternative hypothesis.

Estimating effect size: Cohen's 'd'

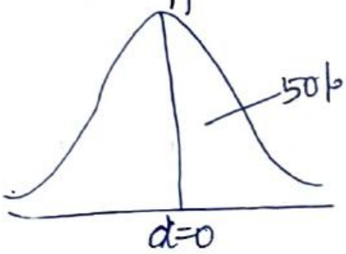
Standardized effect size, Cohen's 'd' (Two independent Samples)

$$d = \frac{\text{mean difference}}{\text{std., deviation}} = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{S_p^2}}$$

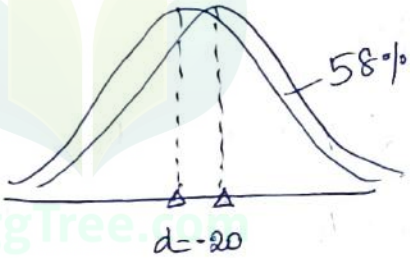
Cohen's Guidelines for 'd'.

d	Effect size	part of SD
·20	Small.	1/5
·50	Medium	1/2.
·80	Large.	4/5

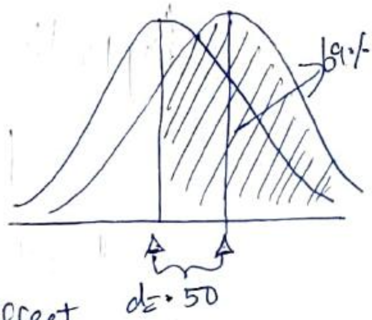
No effect



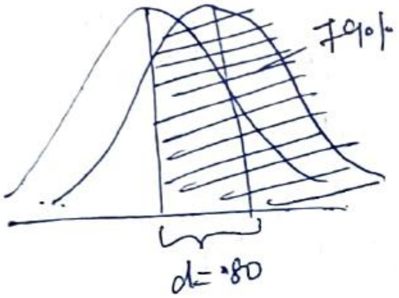
Cohen's small effect



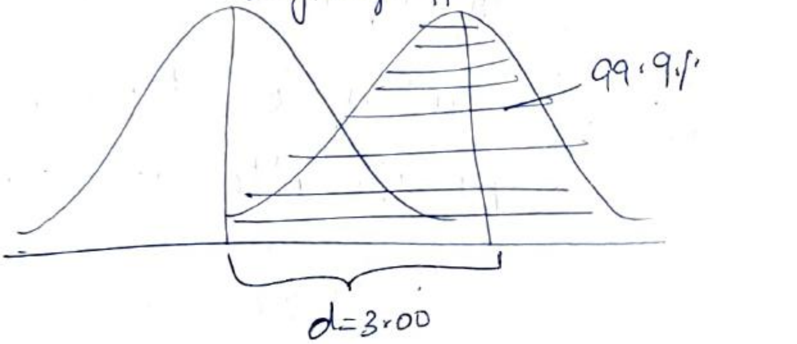
Medium effect.



Cohen's Large effect



Very Large effect



'd' is used to estimate the Standardized effect size for the Statistically Significant results in EPO expt,

$\overline{X_1} - \overline{X_2} = 5$; $SD = 4{\cdot}02$ $= 5/4{\cdot}02 = 1{\cdot}24$ mean difference equivalent to one & one-quarter SD.

widely publicized reports of research findings.

- Ideally, a remedy to the file drawer effect would to be to have all researchers initially register their research project and then report actual data & results of all statistical analyses, whether significant or non significant to a repository of research finds.

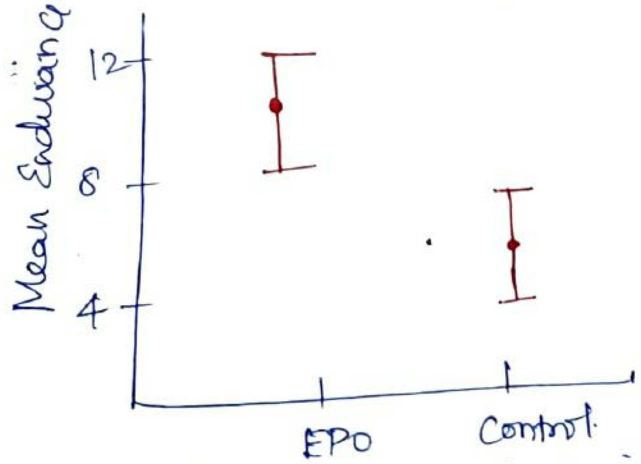
Reports in the Literature.

Parenthetical Statement.

Endurance scores for the EPO group ($\bar{X}=11$, $S=4.24$) significantly exceed those for the control group ($\bar{X}=6$, $S=3.79$) according to a 't' test [$t(10)=2.16$, $p<.05$ & $d=1.24$].

with mean & SD.

The endurance scores for the EPO group ($M=11$, $SD=4.24$) and the ctrl. group ($M=6$, $SD=3.79$) differed significantly [$t(10)=2.16$, $p<.05$ & $d=1.24$]



• → represent the mean endurance.

→ Error bars reflect the stdard error associated with each dot.

→ The non-overlapping error bars imply that differences b/w means might be statistically significant.

Meta-analysis:

- A set of data-collecting and statistical procedures designed to summarize the various effects reported by groups of similar studies. The routine reporting of effect sizes will greatly facilitate efforts to summarize the research findings

Importance of Replication:

- In recent years, we have more health related research findings

- Initially, the hormonal replacement therapy in women decreases the risk of heart attacks and cancer. However, subsequent, more extensive research findings suggested that this therapy has no effect or may even increase these risks.

- One precaution you might adopt is to wait for the replication of any new findings, especially for complex, controversial phenomena.

- There is a well-known bias - often called the "file drawer effect" that favours the publication only of reports that are statistically significant.

- Typically, reports of non-significant findings are never published, but are simply put away in file drawers or waste baskets.

- A solitary significant finding - much like the tip of an iceberg - could be a false positive result reflecting the high cumulative prob. of a type 1 error when there is many unpublicized studies with non-significant findings. This could contribute to the seemingly transitory nature of some

Assumptions:

1) Increase sample size to minimize the effect of any non-normality.

2) Equate sample sizes to minimize the effect of unequal population variances

3) Use a slightly less sensitive, more complex version of 't' designed for unequal variances.

4) Use a less sensitive but more assumption free test such as Mann-Whitney U-test.

t-test for Two Related samples:-

When each subject is measured twice, as in the expt, described, the t test for repeated measures can be extra sensitive to detecting a treatment effect by eliminating the distorting effect of variability due to individual differences

Difference (D) scores:

$$D = X_1 - X_2$$

Where D is the difference score and X_1 and X_2 are the paired endurance scores for each patient measured twice, once under the treatment condn and once under the ctrl., condn respectively.

Mean difference score ($\bar{D}$)

$$\bar{D} = \frac{\Sigma D}{n}$$

where $\bar{D}$ is the mean difference score.

ΣD is (the sum of all +ve difference score) − (the sum of all −ve difference scores)

n is the no of difference scores.

If $\bar{D}$ is +ve ⇒ EPO facilitates endurance.

$\bar{D}$ is −ve ⇒ EPO hinders endurance.

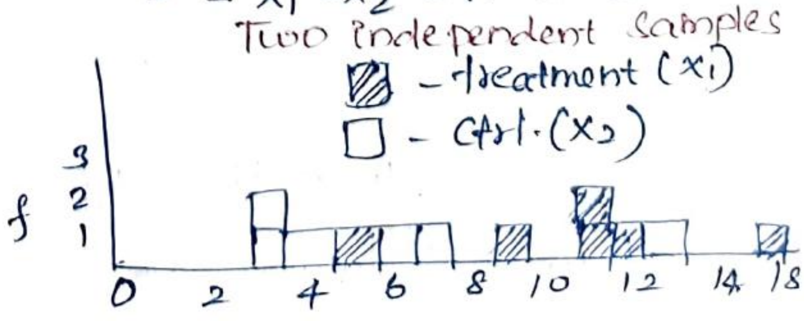
Comparing the two experiments:

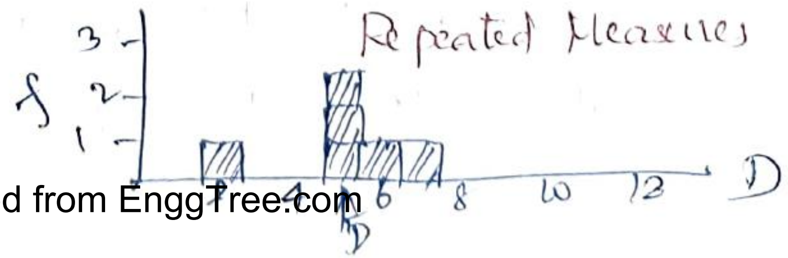
Original		New
X_1	X_2	$D = X_1 - X_2$
12	7	5
5	3	2
11	4	7
11	6	5
9	3	6
18	13	5

$$\bar{X_1} = \frac{\Sigma X_1}{n} = \frac{66}{6} = 11$$

$$\bar{X_2} = \frac{\Sigma X_2}{n_2} = \frac{36}{6} = 6$$

$$\bar{D} = \bar{X_1} - \bar{X_2} = 11 - 6 = 5.$$





From the figure, range of scores in the top histogram of X_1 and X_2 equals 15 (18-3), bottom histogram of D equals only 5 (7-2). This suggests that once the new data have been analyzed with a t-test for repeated measures, it should be possible not only to reject the null-hypothesis again, but also to claim a much smaller p-value than that ($p < .05$) for the t-test for the original expt with two independent samples

Repeated Measure:

- A favorite technique for controlling individual differences is referred to as repeated measures, com each subject is measured more than once.

- By focusing on the differences b/w pairs of scores for each subject, the investigator effectively eliminates by the simple act of subtraction, each individual's unique impact on both endurance scores.

- An analysis of the resulting difference scores reflects only any effects due to EPO, if it exists and random variations of other uncontrolled factors or experimental errors not attributable to individual difference. (Experimental errors refers to random variations in endurance scores due to combined impact of numerous uncontrolled changes, such as slight changes in temperature, treadmill speed etc. as well as any Changes in a particular Subject's motivation, health etc, b/w the two experimental Sessions.

- Coz of smaller standard error term, the result is a test with an increased likelihood of detecting any effect due to EPO.

Two Related Samples:

- Favoured by investigators who wish to control for individual differences, repeated measures represent the most important special case of two related samples.
- 2 related samples occur whenever each observation in one sample is paired, on a one-to-one basis, with a single observation in the other sample.
- An investigator can still choose to use two related samples by matching pairs of different subjects in terms of some uncontrolled variable that appears to have considerable impact on the dependent Variable.
- An investigator still might choose to use two related Samples by matching pairs of different subjects in terms of Some uncontrolled variable that appears to have a Considerable impact on the dependent Variable.

Counterbalancing:

- double duty in both Condn.
- It is customary to randomly assign half of the Subjects to experience the two Condn in a particular

Say first the treatment and then the Ctrl, condn., while the other half of the subjects experience the two condns in the reverse order. It is known as Counterbalancing this adjustment controls for any sequence effect, that is any potential bias in favor of one condn., merely because subjects happen to experience it first

Statistical Hypothesis :

Null hypothesis

Converting to difference score generates a single population of difference scores and the null hypothesis is expressed in terms of this new population. If EPO has either no consistent effect or a -ve effect on endurance scores when patients are measured twice, the population mean of all difference score, μ_D should equal zero or less.

$$H_0 : \mu_D \leq 0$$

Alternative or Research Hypothesis:

The investigator wants to reject the null hypothesis only if EPO actually increases endurance scores.

$H_1 : \mu_D > 0$ (upper tail critical)

$H_1 : \mu_D < 0$ (lower tail critical)

$H_1 : \mu_D \neq 0$ (2 tailed test)

Sampling Distribution of $\bar{D}$.

- The sample mean of the difference scores, $\bar{D}$, varies from sample to sample, and it has a sampling distribution with its own mean, $\mu_{\bar{D}}$ and std, error $\sigma_{\bar{D}}$. When $\bar{D}$ is viewed as the mean for a single sample of difference scores, its sampling distribution can be depicted as a straight forward extension of the sampling distribution of $\bar{X}$, the mean for a single sample of original scores.

- The mean $\mu_{\bar{D}}$, standard error $\sigma_{\bar{D}}$ of the sampling distribution of $\bar{D}$ have essentially the same properties as the mean $\mu_{\bar{X}}$ and std, error $\sigma_{\bar{X}}$ respectively, of the sampling distbn. of $\bar{X}$.

W.K.T $\mu_{\bar{X}} = \mu$., so here. $\mu_{\bar{D}} = \mu_D$

$$\sigma_{\bar{D}} = \frac{\sigma_D}{\sqrt{n}}$$

t-test ratio:

$$t = \frac{(\text{Sample mean difference}) - (\text{hypothesized population mean difference})}{\text{estimated std, error}}$$

$$\boxed{t = \frac{\bar{D} - \mu_{D\,hyp}}{s_{\bar{D}}}}$$

where

- $\bar{D}$ represents the sample means of difference scores.

- $\mu_{D hyp}$ represents the hypothesised population mean for the difference scores;

- $S_{\bar{D}}$ represents the estimated std, error of $\bar{D}$

Finding Critical t-values:

$t_c = 2.015$

Hypothesis Test Summary
t-test for Two population Means:
(Repeated Measures - EPO Experiment)

Research Pblm

When patients are measured twice, once with and once without EPO, does the population mean difference score show greater endurance due to EPO?

Statistical Hypotheses

$H_0: \mu_D \leq 0$

$H_1: \mu_D > 0$

Decision rule:

Reject H_0 at the .05 level of Significance if $t \geq 2.015$

$df = n-1 = 6-1 = 5$

Calculations:

$$t = \frac{5-0}{0.68} = 7.35$$

Decision:

Reject Ho at the .05 level of significance 'Cos the calculated 't0' = 7.35 > 2.015 (t_c)

Interpretation:

There is evidence that when patients are measured twice, EPO is found to increase the mean endurance score.

Calculations for the t-test:-

Panel-I - Sample Std., deviation S_D & the mean

$$SS_D = \Sigma D^2 - \frac{(\Sigma D)^2}{n}$$

$$\bar{D} = \frac{\Sigma D}{n} \; ; \quad S_D = \sqrt{\frac{SS_D}{n-1}}$$ $S_D \to$ Sample std., deviation S_D

Panel-II : Standard error $S_{\bar{D}}$

$$S_{\bar{D}} = \frac{S_D}{\sqrt{n}} \; ;$$

Panel-III : t ratio.

$$t = \frac{\bar{D} - \mu_{D_{hyp}}}{S_{\bar{D}}}$$

Calculations:

	Endurance Scores		Difference Scores	
	EPO	Control		
Patient	X_1	X_2	$D=(X_1-X_2)$	D^2
1	12	7	5	25
2	5	3	2	4
3	11	4	7	49
4	11	6	5	25
5	9	3	6	36
6	18	13	5	25
$n=6$			$\Sigma D=30$	$\Sigma D^2=164$

Panel-I

$$\bar{D} = \frac{\Sigma D}{n} = \frac{30}{6} = 5$$

$$SS_D = \Sigma D^2 - \frac{(\Sigma D)^2}{n} = 164 - \frac{(30)^2}{6} = 164-150=14$$

$$S_D = \sqrt{\frac{SS_D}{n-1}} = \sqrt{\frac{14}{5}} = \sqrt{2.8} = 1.67$$

Panel-II

$$S_{\bar{D}} = \frac{S_D}{\sqrt{n}} = \frac{1.67}{\sqrt{6}} = \frac{1.67}{2.45} = 0.68$$

Panel-III t ratio

$$t \text{ ratio} = \frac{\bar{D} - \mu_{D_{hyp}}}{S_{\bar{D}}} = \frac{5-0}{0.68} = 7.35$$

Summary for EPO Expt :

The boxed hypothesis test summary for the current EPO experiment indicates that ∵ $t_0 = 7.35$ exceeds the $t_c = 2.015$, we reject H_0.

@ $t_0 = 7.35$ with $df = 5$, permits us to claim a much smaller p-value ($P < .001$) thant that ($P < .05$) for a t test based on the same data in the original EPO experiment with two independent samples

Estimating Effect Size :

Confidence Interval for μ_D (2 related Samples)

$$\boxed{\bar{D} \pm (t_{conf})(S_{\bar{D}})}$$

Where $\bar{D}$ represents the sample mean of the differena scores

t_{conf} - Critical t from table

$S_{\bar{D}}$ - estimated std., error.

Finding t_{conf}

Given a value of 2.571 for t_{conf} and values of 5 for $\bar{D}$, the sample mean of the difference scores. and 0.68 for $S_{\bar{D}}$, the estimated std., error

$$\boxed{5 \pm (2.571)(0.68) = 5 \pm 1.75 = \begin{cases} 6.75 \\ 3.25 \end{cases}}$$

It can be claimed with 95% confidence, that the interval b/w 3.25 mins and 6.75 mins includes the true mean for the population of difference endurance scores.

Interpreting Confidence Intervals for μ_D:

- The appearances of the only +ve differences indicates that when patients are measured twice EPO facilitates endurance.
- We can reasonably confident that on average, the true facilitative effect is neither less than 3.25 mins nor more than 6.75 mins

Standardized Effect Size, Cohen's d (Two related samples)

$$\boxed{d = \frac{\overline{D}}{s_D}}$$

where d refers to the standardized estimate of effect size, while $\overline{D}$ and s_D represent the sample mean and std., deviation.

$$\overline{D} = 5 \ ; s_D = 1.67 \ ; \ d = \frac{5}{1.67} = 2.99 \simeq 3$$

Assumptions:

- Whether testing a hypothesis or constructing a confidence interval, 't' assumes that that population of difference scores is normally distributed.

- The unlikely event that you encounter conspicuous departures from normality, consider either increasing the sample size or using the less sensitive but more assumption-free Wilcoxon T test.

Summary of t-Tests for population means

Type of Sample	Sample Mean	Null Hypo	Std. error	tratio	Degrees of freedom
one sample	$\overline{X}$	Ho: μ = some number	$S_{\overline{X}}$	$\frac{\overline{X}-\mu_{hyp}}{S_{\overline{X}}}$	$n-1$
Two indept Samples (no pairing)	$\overline{X}_1-\overline{X}_2$	Ho: $\mu_1-\mu_2=0$	$S_{\overline{X}_1-\overline{X}_2}$	$\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)_{hyp}}{S_{\overline{X}_1-\overline{X}_2}}$	n_1+n_2-2
Two related Samples (pairs of observation	$\overline{D}$	Ho: $\mu_D=0$	$S_{\overline{D}}$	$\frac{\overline{D}-\mu_{D\,hyp}}{S_{\overline{D}}}$	$n-1$

t-test for the population Correlation Coefficient, ρ

A no, b/w +1.00 and -1.00 that describes the linear relationship b/w pairs of quantitative variables for some population.

- Correlation - finding similarity b/w the data.
- 'r' - b/w the no of cards sent and the no of cards received by five friends.

Null hypothesis:

hypo. population Correlation Coefficient - (ρ).

$$t = \frac{r - \rho_{hyp}}{\sqrt{\frac{1-r^2}{n-2}}}$$

r - sample correlation coefficient

n - refers to the no of pairs of observations.

$\sqrt{\frac{1-r^2}{n-2}}$ - estimated std error of the sample correlation coefficient.

Greeting card exchange.

Problem:

Could there be a correlation b/w the no of cards sent and the no of cards received for the population of all friends

Statistical hypothesis:

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

Decision Rule:

Reject Ho at the ·05 level of significance if $t \geq 3.182$ or if $t \leq -3.182$, $df = n-2 = 5-2 = 3$

Calculations:

$r = .80$ & $n = 5$

$$t = \frac{0.80 - 0}{\sqrt{\frac{1-(.80)^2}{5-2}}} = \frac{.80}{\sqrt{\frac{1-.64}{3}}} = \frac{.80}{\sqrt{\frac{.36}{3}}} = \frac{.80}{\sqrt{.12}}$$

$$= \frac{.80}{.35} = \cancel{2.29}\ 2.31$$

Decision:

Retain Ho at the ·05 level of significance 'coz $t = 2.29$ is less positive than 3·182.

Interpretation:

$\rho = 0$; There is no evidence of a relationship b/w the no of cards sent & no of cards received in the population of friends

UNIT V PREDICTIVE ANALYTICS

Linear least squares – implementation – goodness of fit – testing a linear model – weighted Resampling

Linear least squares

Correlation coefficients measure the strength and sign of a relationship, but not the slope. There are several ways to estimate the slope; the most common is a linear least squares fit. A “linear fit” is a line intended to model the relationship between variables. A “least squares” fit is one that minimizes the mean squared error (MSE) between the line and the data.

Suppose we have a sequence of points, ys, that we want to express as a function of another sequence xs. If there is a linear relationship between xs and ys with intercept inter and slope slope, we expect each y[i] to be inter + slope * x[i].

But unless the correlation is perfect, this prediction is only approximate. The vertical deviation from the line, or residual, is res = ys - (inter + slope * xs)

The residuals might be due to random factors like measurement error, or nonrandom factors that are unknown. For example, if we are trying to predict weight as a function of height, unknown factors might include diet, exercise, and body type.

If we get the parameters inter and slope wrong, the residuals get bigger, so it makes intuitive sense that the parameters we want are the ones that minimize the residuals.

We might try to minimize the absolute value of the residuals, or their squares, or their cubes; but the most common choice is to minimize the sum of squared residuals, sum(res**2).

Why? There are three good reasons and one less important one:

- Squaring has the feature of treating positive and negative residuals the same, which is usually what we want.

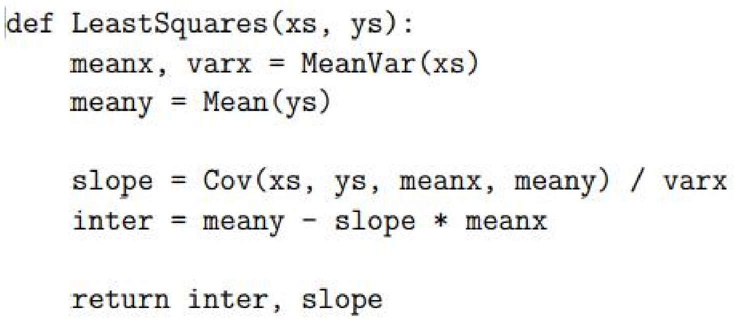
- Squaring gives more weight to large residuals, but not so much weight that the largest residual always dominates.

- If the residuals are uncorrelated and normally distributed with mean 0 and constant (but unknown) variance, then the least squares fit is also the maximum likelihood estimator of inter and slope.

- The values of inter and slope that minimize the squared residuals can be computed efficiently.

Implementation

thinkstats2 provides simple functions that demonstrate linear least squares:



```
def LeastSquares(xs, ys):
    meanx, varx = MeanVar(xs)
    meany = Mean(ys)

    slope = Cov(xs, ys, meanx, meany) / varx
    inter = meany - slope * meanx

    return inter, slope
```

LeastSquares takes sequences xs and ys and returns the estimated parameters inter and slope. thinkstats2 also provides FitLine, which takes inter and slope and returns the fitted line for a sequence of xs.

```
def FitLine(xs, inter, slope):
    fit_xs = np.sort(xs)
    fit_ys = inter + slope * fit_xs
    return fit_xs, fit_ys
```

We can use these functions to compute the least squares fit for birth weight as a function of mother's age.

```
    live, firsts, others = first.MakeFrames()
    live = live.dropna(subset=['agepreg', 'totalwgt_lb'])
    ages = live.agepreg
    weights = live.totalwgt_lb

    inter, slope = thinkstats2.LeastSquares(ages, weights)
    fit_xs, fit_ys = thinkstats2.FitLine(ages, inter, slope)
```

The estimated intercept and slope are 6.8 lbs and 0.017 lbs per year. These values are hard to interpret in this form: the intercept is the expected weight of a baby whose mother is 0 years old, which doesn't make sense in context, and the slope is too small to grasp easily.

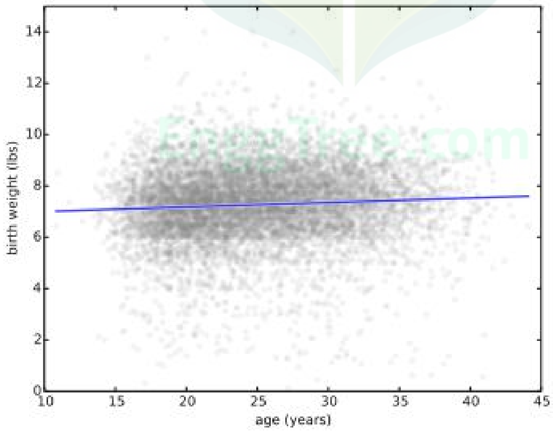
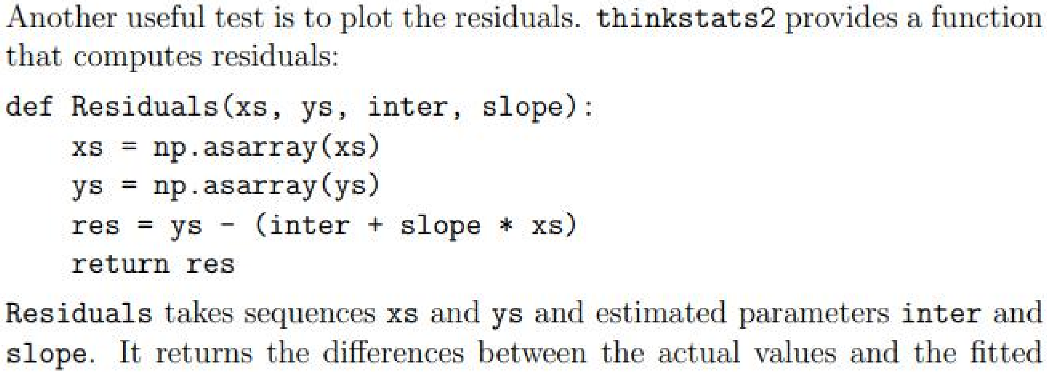


Figure 10.1: Scatter plot of birth weight and mother's age with a linear fit.

Instead of presenting the intercept at $x = 0$, it is often helpful to present the intercept at the mean of x. In this case the mean age is about 25 years and the mean baby weight for a 25 year old mother is 7.3 pounds. The slope is 0.27 ounces per year, or 0.17 pounds per decade.

Figure 10.1 shows a scatter plot of birth weight and age along with the fitted line. It's a good idea to look at a figure like this to assess whether the relationship is linear and whether the fitted line seems like a good model of the relationship.

Residuals



Another useful test is to plot the residuals. thinkstats2 provides a function that computes residuals:

```
def Residuals(xs, ys, inter, slope):
    xs = np.asarray(xs)
    ys = np.asarray(ys)
    res = ys - (inter + slope * xs)
    return res
```

Residuals takes sequences xs and ys and estimated parameters inter and slope. It returns the differences between the actual values and the fitted

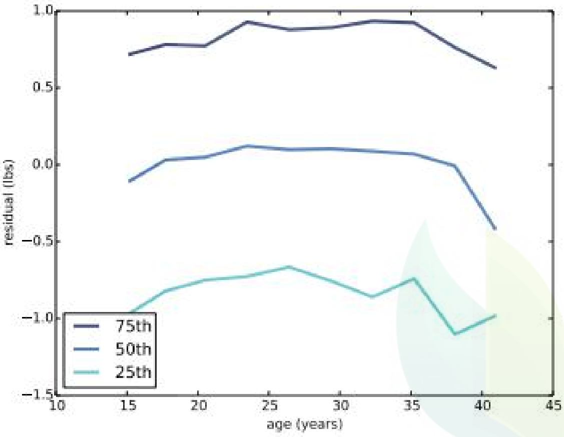


Figure 10.2: Residuals of the linear fit.

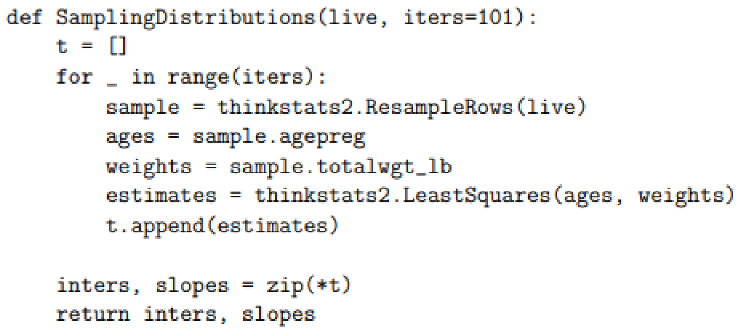
To visualize the residuals, I group respondents by age and compute percentiles in each group, as we saw in Section 7.2. Figure 10.2 shows the 25th,50th and 75th percentiles of the residuals for each age group. The median is near zero, as expected, and the interquartile range is about 2 pounds. So if we know the mother's age, we can guess the baby's weight within a pound,about 50% of the time.

Ideally these lines should be flat, indicating that the residuals are random,and parallel, indicating that the variance of the residuals is the same for all age groups. In fact, the lines are close to parallel, so that's good; but they have some curvature, indicating that the relationship is nonlinear. Nevertheless, the linear fit is a simple model that is probably good enough for some purposes.

Estimation

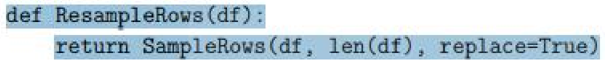
The parameters slope and inter are estimates based on a sample; like other estimates, they are vulnerable to sampling bias, measurement error, and sampling error. As discussed in Chapter 8, sampling bias is caused by non-representative sampling, measurement error is caused by errors in collecting and recording data, and sampling error is the result of measuring a sample rather than the entire population.

I simulate the experiments by resampling the data; that is, I treat the observed pregnancies as if they were the entire population and draw samples, with replacement, from the observed sample.

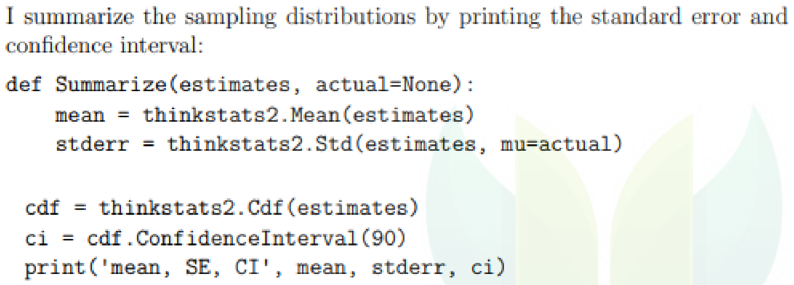


```
def SamplingDistributions(live, iters=101):
    t = []
    for _ in range(iters):
        sample = thinkstats2.ResampleRows(live)
        ages = sample.agepreg
        weights = sample.totalwgt_lb
        estimates = thinkstats2.LeastSquares(ages, weights)
        t.append(estimates)

    inters, slopes = zip(*t)
    return inters, slopes
```

SamplingDistributions takes a DataFrame with one row per live birth, and iters, the number of experiments to simulate. It uses ResampleRows to resample the observed pregnancies. We've already seen SampleRows, which chooses random rows from a DataFrame. thinkstats2 also provides ResampleRows, which returns a sample the same size as the original:



```
def ResampleRows(df):
    return SampleRows(df, len(df), replace=True)
```



I summarize the sampling distributions by printing the standard error and confidence interval:

```
def Summarize(estimates, actual=None):
    mean = thinkstats2.Mean(estimates)
    stderr = thinkstats2.Std(estimates, mu=actual)

  cdf = thinkstats2.Cdf(estimates)
  ci = cdf.ConfidenceInterval(90)
  print('mean, SE, CI', mean, stderr, ci)
```

Summarize takes a sequence of estimates and the actual value. It prints the mean of the estimates, the standard error and a 90% confidence interval. For the intercept, the mean estimate is 6.83, with standard error 0.07 and 90% confidence interval (6.71, 6.94). The estimated slope, in more compact form, is 0.0174, SE 0.0028, CI (0.0126, 0.0220). There is almost a factor of two between the low and high ends of this CI, so it should be considered a rough estimate.

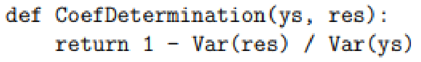
Goodness of fit:

There are several ways to measure the quality of a linear model, or goodness of fit. One of the simplest is the standard deviation of the residuals.

If you use a linear model to make predictions, Std(res) is the root mean squared error (RMSE) of your predictions. For example, if you use mother's age to guess birth weight, the RMSE of your guess would be 1.40 lbs.

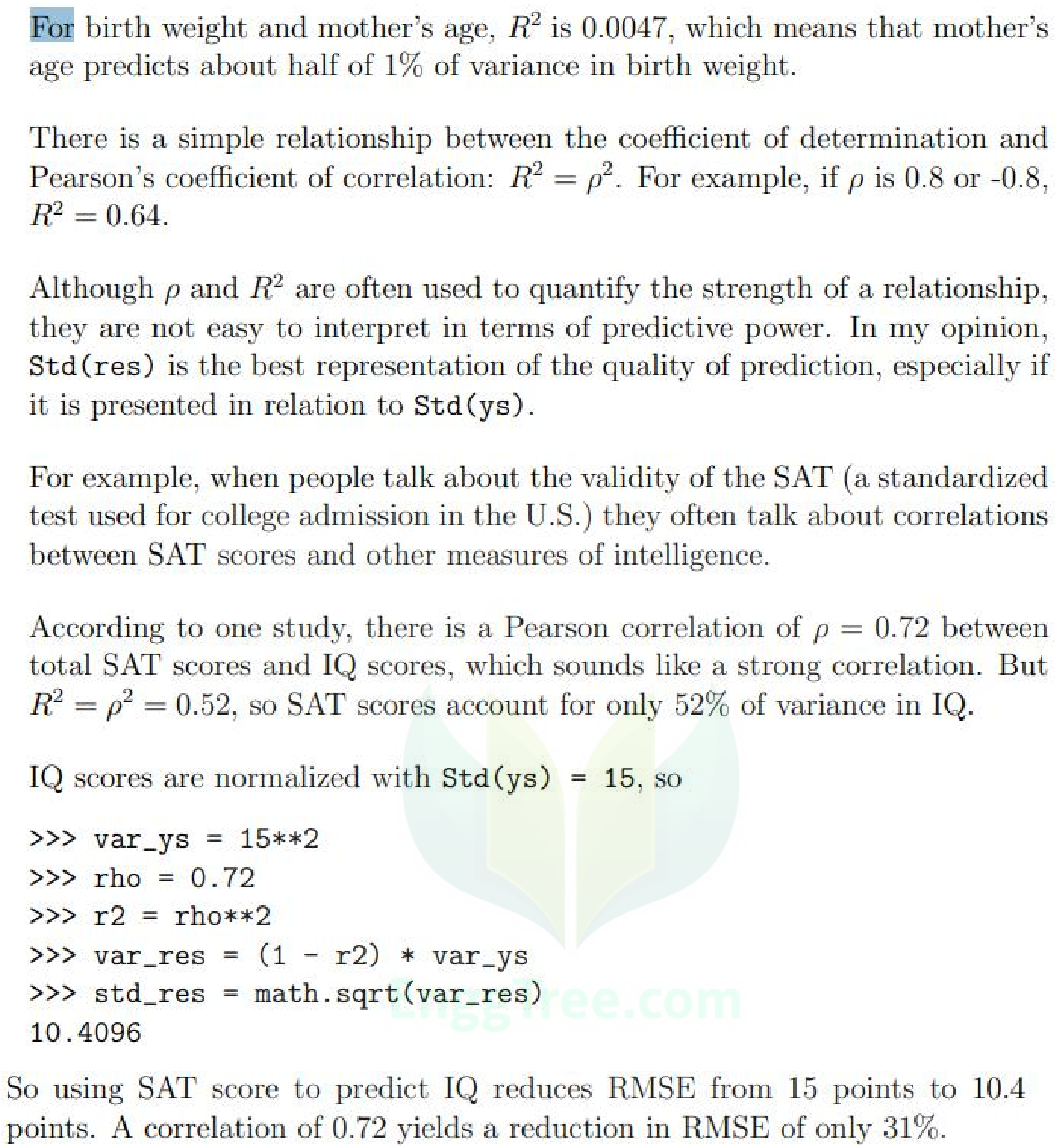
If you guess birth weight without knowing the mother's age, the RMSE of your guess is Std(ys), which is 1.41 lbs. So in this example, knowing a mother's age does not improve the predictions substantially.

Another way to measure goodness of fit is the coefficient of determination, usually denoted R2 and called "R-squared":



```
def CoefDetermination(ys, res):
    return 1 - Var(res) / Var(ys)
```

Var(res) is the MSE of your guesses using the model, Var(ys) is the MSE without it. So their ratio is the fraction of MSE that remains if you use the model, and R^2 is the fraction of MSE the model eliminates.



For birth weight and mother's age, R^2 is 0.0047, which means that mother's age predicts about half of 1% of variance in birth weight.

There is a simple relationship between the coefficient of determination and Pearson's coefficient of correlation: $R^2 = \rho^2$. For example, if ρ is 0.8 or -0.8, $R^2 = 0.64$.

Although ρ and R^2 are often used to quantify the strength of a relationship, they are not easy to interpret in terms of predictive power. In my opinion, `Std(res)` is the best representation of the quality of prediction, especially if it is presented in relation to `Std(ys)`.

For example, when people talk about the validity of the SAT (a standardized test used for college admission in the U.S.) they often talk about correlations between SAT scores and other measures of intelligence.

According to one study, there is a Pearson correlation of $\rho = 0.72$ between total SAT scores and IQ scores, which sounds like a strong correlation. But $R^2 = \rho^2 = 0.52$, so SAT scores account for only 52% of variance in IQ.

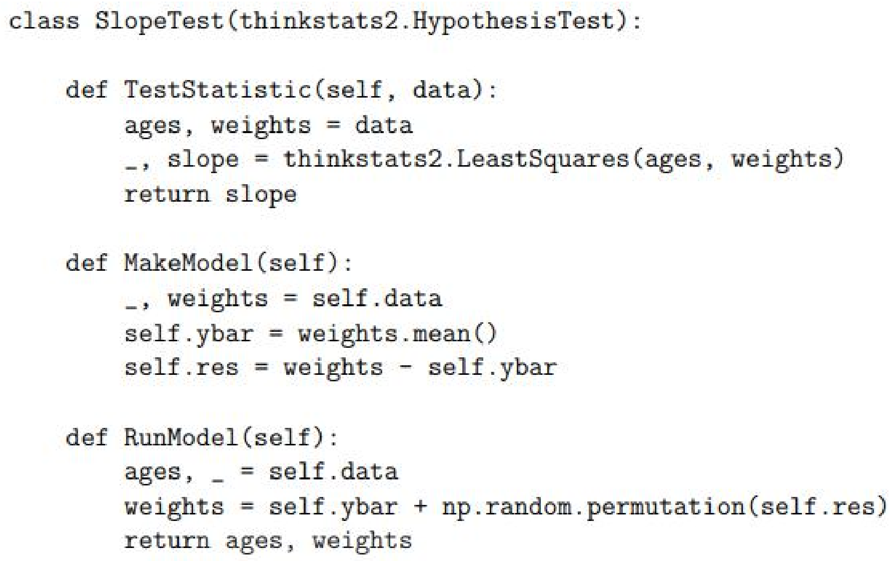
IQ scores are normalized with `Std(ys) = 15`, so

```
>>> var_ys = 15**2
>>> rho = 0.72
>>> r2 = rho**2
>>> var_res = (1 - r2) * var_ys
>>> std_res = math.sqrt(var_res)
10.4096
```

So using SAT score to predict IQ reduces RMSE from 15 points to 10.4 points. A correlation of 0.72 yields a reduction in RMSE of only 31%.

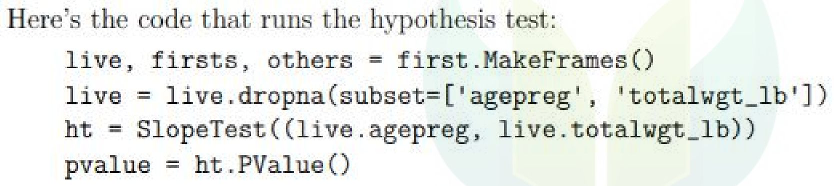
Testing a linear model

The effect of mother's age on birth weight is small, and has little predictive power. So is it possible that the apparent relationship is due to chance? There are several ways we might test the results of a linear fit. One option is to test whether the apparent reduction in MSE is due to chance. In that case, the test statistic is R^2 and the null hypothesis is that there is no relationship between the variables. We can simulate the null hypothesis by permutation, as in Section 9.5, when we tested the correlation between mother's age and birth weight. In fact, because $R^2 = \rho$ 2 , a one-sided test of R2 is equivalent to a two-sided test of ρ. We've already done that test, and found $p < 0.001$, so we conclude that the apparent relationship between mother's age and birth weight is statistically significant. Another approach is to test whether the apparent slope is due to chance. The null hypothesis is that the slope is actually zero; in that case we can model the birth weights as random variations around their mean. Here's a HypothesisTest for this model



```
class SlopeTest(thinkstats2.HypothesisTest):

    def TestStatistic(self, data):
        ages, weights = data
        _, slope = thinkstats2.LeastSquares(ages, weights)
        return slope

    def MakeModel(self):
        _, weights = self.data
        self.ybar = weights.mean()
        self.res = weights - self.ybar

    def RunModel(self):
        ages, _ = self.data
        weights = self.ybar + np.random.permutation(self.res)
        return ages, weights
```

The data are represented as sequences of ages and weights. The test statistic is the slope estimated by LeastSquares. The model of the null hypothesis is represented by the mean weight of all babies and the deviations from the mean. To generate simulated data, we permute the deviations and add them to the mean.



Here's the code that runs the hypothesis test:

```
live, firsts, others = first.MakeFrames()
live = live.dropna(subset=['agepreg', 'totalwgt_lb'])
ht = SlopeTest((live.agepreg, live.totalwgt_lb))
pvalue = ht.PValue()
```

The p-value is less than 0.001, so although the estimated slope is small, it is unlikely to be due to chance. Estimating the p-value by simulating the null hypothesis is strictly correct, but there is a simpler alternative. Remember that we already computed the sampling distribution of the slope. The sampling distribution is centered about the estimated slope, 0.017 lbs/year, and the slopes under the null hypothesis are centered around 0; but other than that, the distributions are identical.

So we could estimate the p-value two ways:

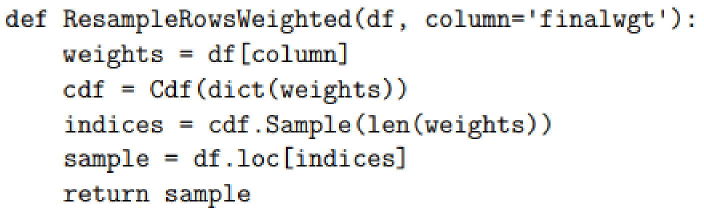
- Compute the probability that the slope under the null hypothesis exceeds the observed slope.

- Compute the probability that the slope in the sampling distribution falls below 0. (If the estimated slope were negative, we would compute the probability that the slope in the sampling distribution exceeds 0.)

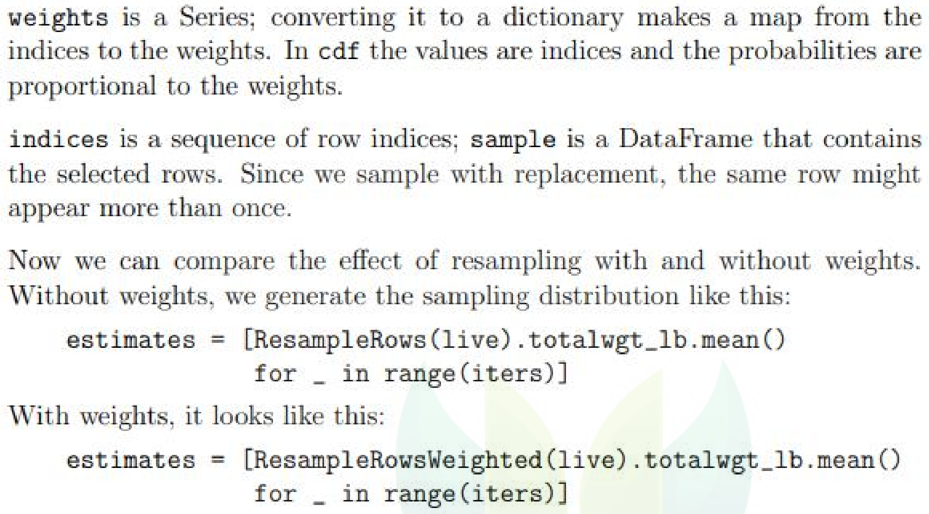
The second option is easier because we normally want to compute the sampling distribution of the parameters anyway. And it is a good approximation unless the sample size is small and the distribution of residuals is skewed. Even then, it is usually good enough, because p-values don't have to be precise.

Weighted resampling

As an example, if you survey 100,000 people in a country of 300 million, each respondent represents 3,000 people. If you oversample one group by a factor of 2, each person in the oversampled group would have a lower weight, about 1500. To correct for oversampling, we can use resampling; that is, we can draw samples from the survey using probabilities proportional to sampling weights. Then, for any quantity we want to estimate, we can generate sampling distributions, standard errors, and confidence intervals. As an example, I will estimate mean birth weight with and without sampling weights. we saw ResampleRows, which chooses rows from a DataFrame, giving each row the same probability. Now we need to do the same thing using probabilities proportional to sampling weights. ResampleRowsWeighted takes a DataFrame, resamples rows according to the weights in finalwgt, and returns a DataFrame containing the resampled rows:



```
def ResampleRowsWeighted(df, column='finalwgt'):
    weights = df[column]
    cdf = Cdf(dict(weights))
    indices = cdf.Sample(len(weights))
    sample = df.loc[indices]
    return sample
```



weights is a Series; converting it to a dictionary makes a map from the indices to the weights. In cdf the values are indices and the probabilities are proportional to the weights.

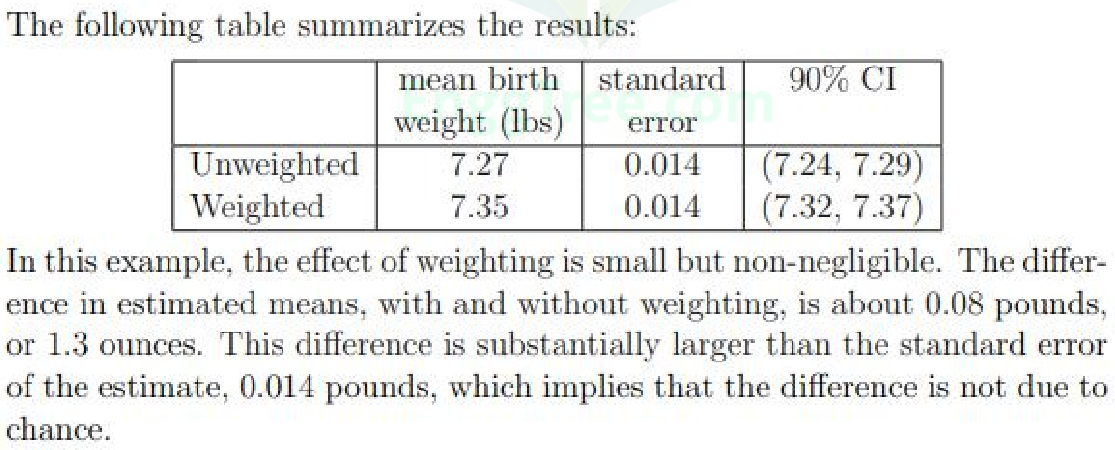
indices is a sequence of row indices; sample is a DataFrame that contains the selected rows. Since we sample with replacement, the same row might appear more than once.

Now we can compare the effect of resampling with and without weights. Without weights, we generate the sampling distribution like this:

```
    estimates = [ResampleRows(live).totalwgt_lb.mean()
                 for _ in range(iters)]
```

With weights, it looks like this:

```
    estimates = [ResampleRowsWeighted(live).totalwgt_lb.mean()
                 for _ in range(iters)]
```



The following table summarizes the results:

	mean birth weight (lbs)	standard error	90% CI
Unweighted	7.27	0.014	(7.24, 7.29)
Weighted	7.35	0.014	(7.32, 7.37)

In this example, the effect of weighting is small but non-negligible. The difference in estimated means, with and without weighting, is about 0.08 pounds, or 1.3 ounces. This difference is substantially larger than the standard error of the estimate, 0.014 pounds, which implies that the difference is not due to chance.

Regression using StatsModels – multiple regression – nonlinear relationships – logistic regression– estimating parameters – accuracy

Regression

The linear least squares fit in the previous chapter is an example of regression, which is the more general problem of fitting any kind of model to any kind of data. This use of the term "regression" is a historical accident; it is only indirectly related to the original meaning of the word.

The goal of regression analysis is to describe the relationship between one set of variables, called the dependent variables, and another set of variables, called independent or explanatory variables.

We used mother's age as an explanatory variable to predict birth weight as a dependent variable. When there is only one dependent and one explanatory variable, that's simple regression. In this chapter, we move on to multiple regression, with more than one explanatory variable. If there is more than one dependent variable, that's multivariate regression.

If the relationship between the dependent and explanatory variable is linear, that's linear regression. For example, if the dependent variable is y and the explanatory variables are x1 and x2, we would write the following linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where β_0 is the intercept, β_1 is the parameter associated with x_1, β_2 is the parameter associated with x_2, and ε is the residual due to random variation or other unknown factors.

Given a sequence of values for y and sequences for x1 and x2, we can find the parameters, β_0, β_1, and β_2, that minimize the sum of ε^2 . This process is called ordinary least squares.

StatsModels

```
import statsmodels.formula.api as smf

live, firsts, others = first.MakeFrames()
formula = 'totalwgt_lb ~ agepreg'
model = smf.ols(formula, data=live)
results = model.fit()
```

statsmodels provides two interfaces (APIs); the "formula" API uses strings to identify the dependent and explanatory variables. It uses a syntax called patsy; in this example, the ~ operator separates the dependent variable on the left from the explanatory variables on the right.

smf.ols takes the formula string and the DataFrame, live, and returns an OLS object that represents the model. The name ols stands for "ordinary least squares."

The fit method fits the model to the data and returns a RegressionResults object that contains the results.

The results are also available as attributes. params is a Series that maps from variable names to their parameters, so we can get the intercept and slope like this:

inter = results.params['Intercept'] slope = results.params['agepreg']

The estimated parameters are 6.83 and 0.0175, the same as from LeastSquares.

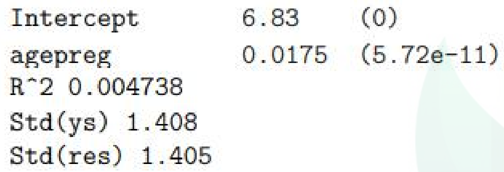
pvalues is a Series that maps from variable names to the associated p-values, so we can check whether the estimated slope is statistically significant:

slope_pvalue = results.pvalues['agepreg']

The p-value associated with agepreg is 5.7e-11, which is less than 0.001, as expected.

results.rsquared contains R^2 , which is 0.0047. results also provides f_pvalue, which is the p-value associated with the model as a whole, similar to testing whether R2 is statistically significant. And results provides resid, a sequence of residuals, and fittedvalues, a sequence of fitted values corresponding to agepreg.

Here are the results of this model:



```
Intercept       6.83    (0)
agepreg         0.0175  (5.72e-11)
R^2 0.004738
Std(ys) 1.408
Std(res) 1.405
```

Std(ys) is the standard deviation of the dependent variable, which is the RMSE if you have to guess birth weights without the benefit of any explanatory variables. Std(res) is the standard deviation of the residuals, which is the RMSE if your guesses are informed by the mother’s age. As we have already seen, knowing the mother’s age provides no substantial improvement to the predictions.

Multiple regression

In Section 4.5 we saw that first babies tend to be lighter than others, and this effect is statistically significant. But it is a strange result because there is no obvious mechanism that would cause first babies to be lighter. So we might wonder whether this relationship is spurious.

In fact, there is a possible explanation for this effect. We have seen that birth weight depends on mother’s age, and we might expect that mothers of first babies are younger than others.

With a few calculations we can check whether this explanation is plausible.

Then we’ll use multiple regression to investigate more carefully. First, let’s see how big the difference in weight is:

diff_weight = firsts.totalwgt_lb.mean() - others.totalwgt_lb.mean()

First babies are 0.125 lbs lighter, or 2 ounces. And the difference in ages:

diff_age = firsts.agepreg.mean() - others.agepreg.mean()

The mothers of first babies are 3.59 years younger. Running the linear model again, we get the change in birth weight as a function of age:

```
results = smf.ols('totalwgt_lb ~ agepreg', data=live).fit()

slope = results.params['agepreg']
```

The slope is 0.0175 pounds per year. If we multiply the slope by the difference in ages, we get the expected difference in birth weight for first babies and

others, due to mother’s age:

```
slope * diff_age
```

The result is 0.063, just about half of the observed difference. So we conclude, tentatively, that the observed difference in birth weight can be partly explained by the difference in mother’s age.

Using multiple regression, we can explore these relationships more systematically.

```
live['isfirst'] = live.birthord == 1

formula = 'totalwgt_lb ~ isfirst'

results = smf.ols(formula, data=live).fit()
```

The first line creates a new column named isfirst that is True for first

babies and false otherwise. Then we fit a model using is first as an explanatory variable.

Here are the results:

```
Intercept 7.33 (0)

isfirst[T.True] -0.125 (2.55e-05)

R^2 0.00196
```

Because isfirst is a boolean, ols treats it as a categorical variable, which means that the values fall into categories, like True and False, and should not be treated as numbers. The estimated parameter is the effect on birth weight when isfirst is true, so the result, -0.125 lbs, is the difference in birth weight between first babies and others.

The slope and the intercept are statistically significant, which means that they were unlikely to occur by chance , but the the R2 value for this model is small, which means that isfirst doesn’t account for a substantial part of the variation in birth weight.

The results are similar with agepreg:

```
Intercept 6.83 (0)

agepreg 0.0175 (5.72e-11)

R^2 0.004738
```

Again, the parameters are statistically significant, but R^2 is low.

These models confirm results we have already seen. But now we can fit a single model that includes both variables. With the formula

totalwgt_lb ~ isfirst + agepreg, we get:

Intercept 6.91 (0)

isfirst[T.True] -0.0698 (0.0253)

agepreg 0.0154 (3.93e-08)

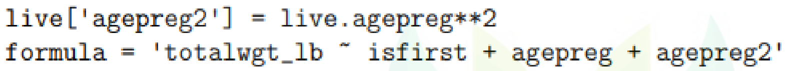
R^2 0.005289

In the combined model, the parameter for isfirst is smaller by about half, which means that part of the apparent effect of isfirst is actually accounted for by agepreg. And the p-value for isfirst is about 2.5%, which is on the border of statistical significance.

R^2 for this model is a little higher, which indicates that the two variables together account for more variation in birth weight than either alone (but not by much).

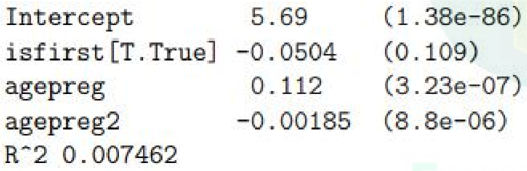
Nonlinear relationships

Remembering that the contribution of agepreg might be nonlinear, we might consider adding a variable to capture more of this relationship. One option is to create a column, agepreg2, that contains the squares of the ages:



```
live['agepreg2'] = live.agepreg**2
formula = 'totalwgt_lb ~ isfirst + agepreg + agepreg2'
```

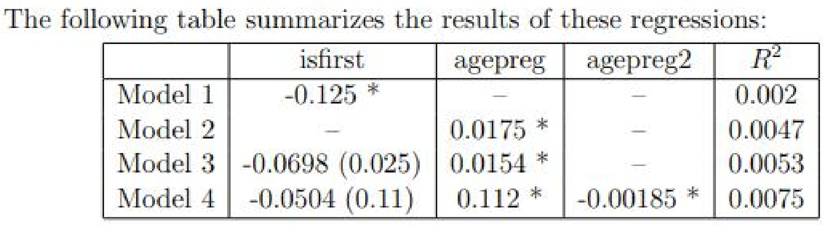
Now by estimating parameters for agepreg and agepreg2, we are effectively fitting a parabola:



```
Intercept         5.69      (1.38e-86)
isfirst[T.True] -0.0504     (0.109)
agepreg           0.112     (3.23e-07)
agepreg2        -0.00185    (8.8e-06)
R^2 0.007462
```

The parameter of agepreg2 is negative, so the parabola curves downward, which is consistent with the shape of the lines. The quadratic model of agepreg accounts for more of the variability in birth weight; the parameter for is first is smaller in this model, and no longer statistically significant.

Using computed variables like agepreg2 is a common way to fit polynomials and other functions to data. This process is still considered linear regression, because the dependent variable is a linear function of the explanatory variables, regardless of whether some variables are nonlinear functions of others.



The following table summarizes the results of these regressions:

	isfirst	agepreg	agepreg2	R^2
Model 1	-0.125 *	–	–	0.002
Model 2	–	0.0175 *	–	0.0047
Model 3	-0.0698 (0.025)	0.0154 *	–	0.0053
Model 4	-0.0504 (0.11)	0.112 *	-0.00185 *	0.0075

The columns in this table are the explanatory variables and the coefficient of determination, R2. Each entry is an estimated parameter and either a p-value in parentheses or an asterisk to indicate a p-value less that 0.001. We conclude that the apparent difference in birth weight is explained, atleast in part, by the difference in mother's age. When we include mother's age in the model, the effect of isfirst gets smaller, and the remaining effect might be due to chance.

In this example, mother's age acts as a control variable; including agepreg in the model "controls for" the difference in age between first-time mothers and others, making it possible to isolate the effect (if any) of isfirst.

Logistic regression

Linear regression can be generalized to handle other kinds of dependent variables. If the dependent variable is boolean, the generalized model is called logistic regression. If the dependent variable is an integer count, it's called

Poisson regression.

As an example of logistic regression, let's consider a variation on the office pool scenario. Suppose a friend of yours is pregnant and you want to predict whether the baby is a boy or a girl. You could use data from the NSFG to find factors that affect the "sex ratio", which is conventionally defined to be the probability of having a boy.

If you encode the dependent variable numerically, for example 0 for a girl and 1 for a boy, you could apply ordinary least squares, but there would be problems. The linear model might be something like this:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Where y is the dependent variable, and x1 and x2 are explanatory variables. Then we could find the parameters that minimize the residuals. The problem with this approach is that it produces predictions that are hard to interpret. Given estimated parameters and values for x1 and x2, the model might predict y = 0.5, but the only meaningful values of y are 0 and 1.

It is tempting to interpret a result like that as a probability; for example,we might say that a respondent with particular values of x1 and x2 has a 50% chance of having a boy. But it is also possible for this model to predict

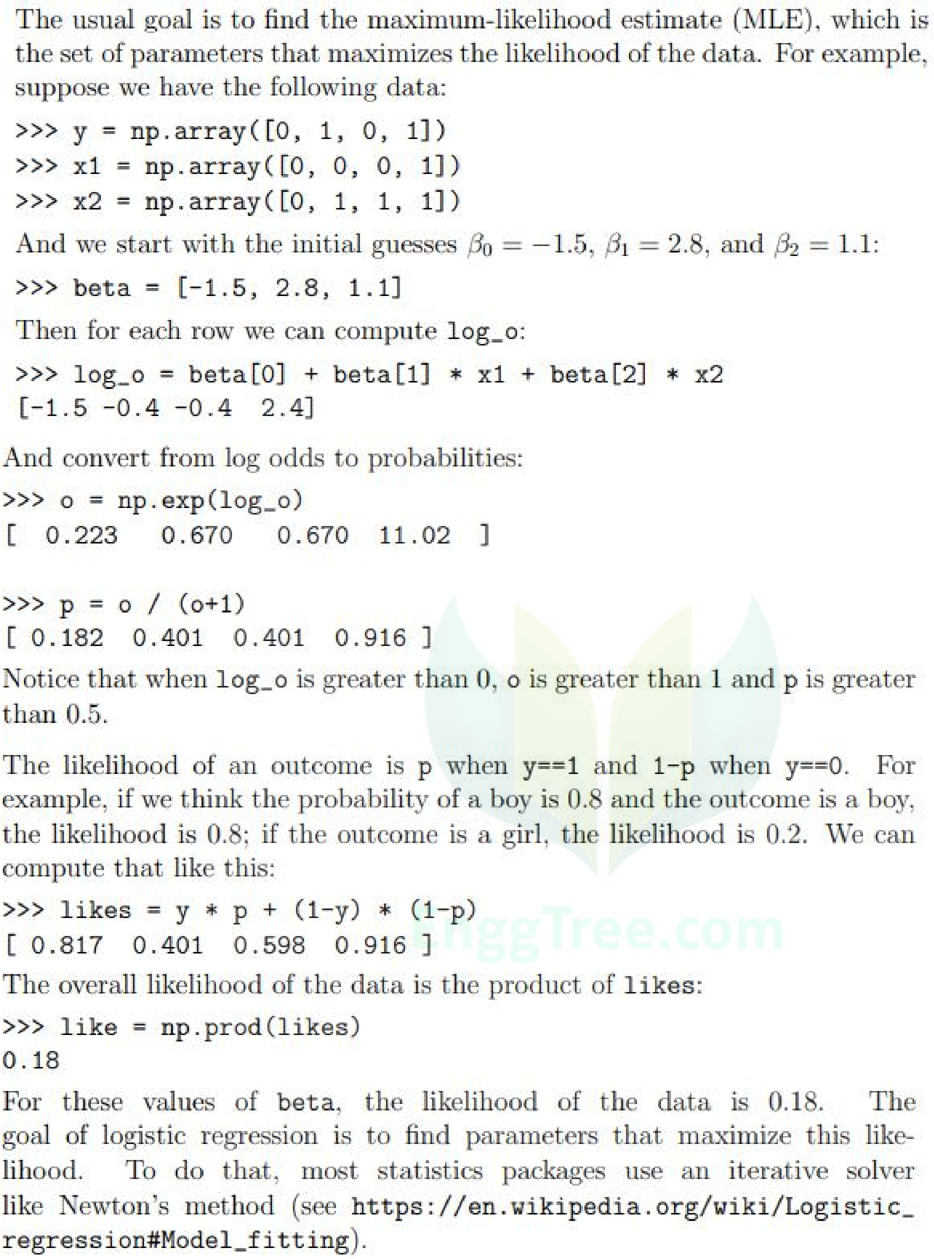
y = 1.1 or y = −0.1, and those are not valid probabilities.

Logistic regression avoids this problem by expressing predictions in terms of odds rather than probabilities. If you are not familiar with odds, "odds in favor" of an event is the ratio of the probability it will occur to the probability that it will not.

Logistic regression is based on the following model:

$$\log o = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Estimating parameters



The usual goal is to find the maximum-likelihood estimate (MLE), which is the set of parameters that maximizes the likelihood of the data. For example, suppose we have the following data:

```
>>> y = np.array([0, 1, 0, 1])
>>> x1 = np.array([0, 0, 0, 1])
>>> x2 = np.array([0, 1, 1, 1])
```

And we start with the initial guesses $\beta_0 = -1.5$, $\beta_1 = 2.8$, and $\beta_2 = 1.1$:

```
>>> beta = [-1.5, 2.8, 1.1]
```

Then for each row we can compute log_o:

```
>>> log_o = beta[0] + beta[1] * x1 + beta[2] * x2
[-1.5 -0.4 -0.4  2.4]
```

And convert from log odds to probabilities:

```
>>> o = np.exp(log_o)
[  0.223   0.670   0.670  11.02  ]

>>> p = o / (o+1)
[ 0.182  0.401  0.401  0.916 ]
```

Notice that when log_o is greater than 0, o is greater than 1 and p is greater than 0.5.

The likelihood of an outcome is p when y==1 and 1-p when y==0. For example, if we think the probability of a boy is 0.8 and the outcome is a boy, the likelihood is 0.8; if the outcome is a girl, the likelihood is 0.2. We can compute that like this:

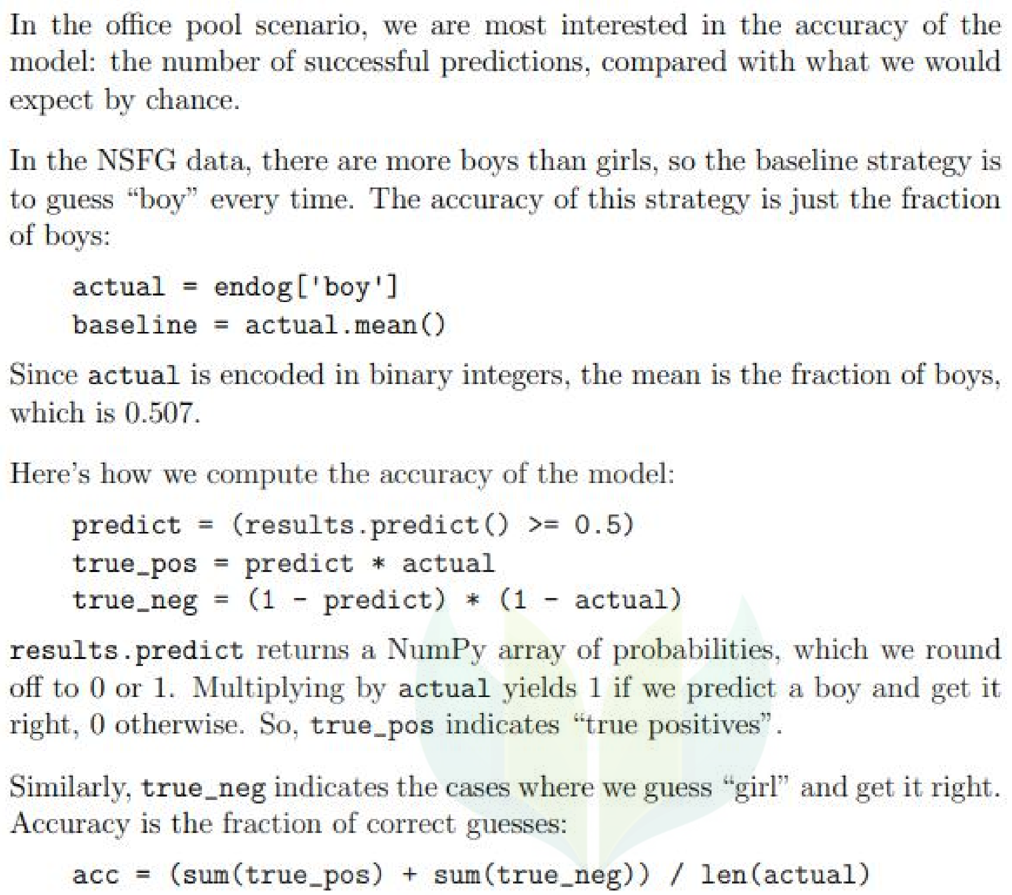
```
>>> likes = y * p + (1-y) * (1-p)
[ 0.817  0.401  0.598  0.916 ]
```

The overall likelihood of the data is the product of likes:

```
>>> like = np.prod(likes)
0.18
```

For these values of beta, the likelihood of the data is 0.18. The goal of logistic regression is to find parameters that maximize this likelihood. To do that, most statistics packages use an iterative solver like Newton's method (see https://en.wikipedia.org/wiki/Logistic_regression#Model_fitting).

Accuracy



In the office pool scenario, we are most interested in the accuracy of the model: the number of successful predictions, compared with what we would expect by chance.

In the NSFG data, there are more boys than girls, so the baseline strategy is to guess “boy” every time. The accuracy of this strategy is just the fraction of boys:

```
    actual = endog['boy']
    baseline = actual.mean()
```

Since `actual` is encoded in binary integers, the mean is the fraction of boys, which is 0.507.

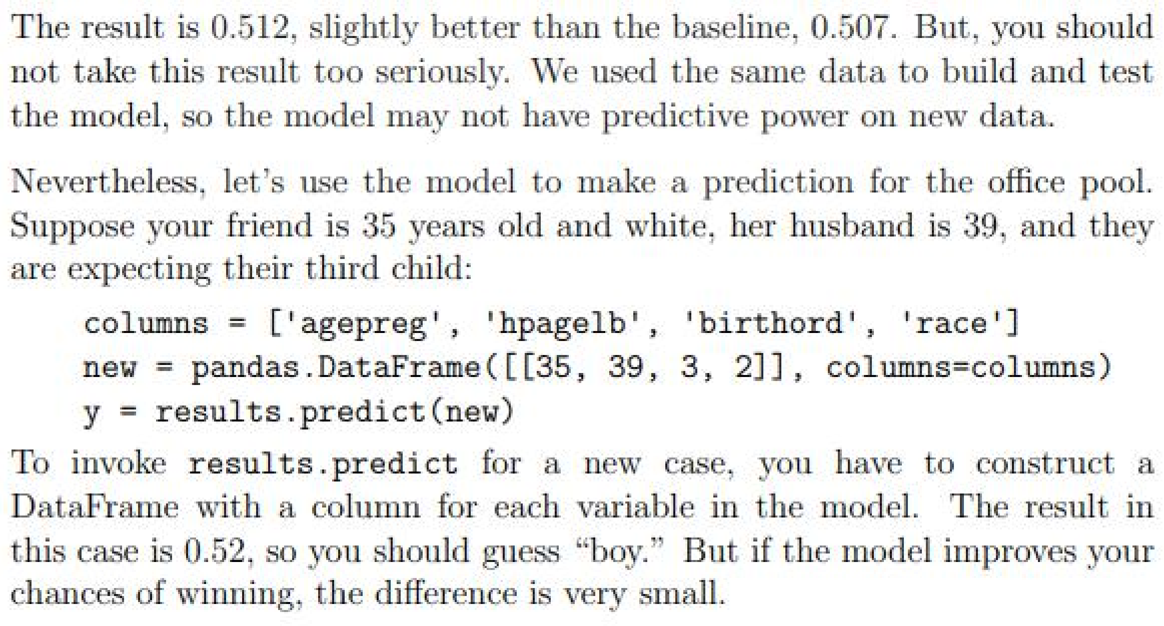
Here’s how we compute the accuracy of the model:

```
    predict = (results.predict() >= 0.5)
    true_pos = predict * actual
    true_neg = (1 - predict) * (1 - actual)
```

`results.predict` returns a NumPy array of probabilities, which we round off to 0 or 1. Multiplying by `actual` yields 1 if we predict a boy and get it right, 0 otherwise. So, `true_pos` indicates “true positives”.

Similarly, `true_neg` indicates the cases where we guess “girl” and get it right. Accuracy is the fraction of correct guesses:

```
    acc = (sum(true_pos) + sum(true_neg)) / len(actual)
```



The result is 0.512, slightly better than the baseline, 0.507. But, you should not take this result too seriously. We used the same data to build and test the model, so the model may not have predictive power on new data.

Nevertheless, let’s use the model to make a prediction for the office pool. Suppose your friend is 35 years old and white, her husband is 39, and they are expecting their third child:

```
    columns = ['agepreg', 'hpagelb', 'birthord', 'race']
    new = pandas.DataFrame([[35, 39, 3, 2]], columns=columns)
    y = results.predict(new)
```

To invoke `results.predict` for a new case, you have to construct a DataFrame with a column for each variable in the model. The result in this case is 0.52, so you should guess “boy.” But if the model improves your chances of winning, the difference is very small.

Time series analysis – moving averages – missing values – serial correlation – autocorrelation Introduction to survival analysis

Time series analysis

A time series is a sequence of measurements from a system that varies in time. One famous example is the “hockey stick graph” that shows global average temperature over time .

Moving averages

Most time series analysis is based on the modeling assumption that the observed series is the sum of three components:

• Trend: A smooth function that captures persistent changes.

• Seasonality: Periodic variation, possibly including daily, weekly,

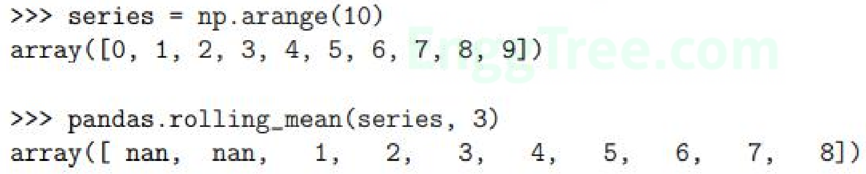
monthly, or yearly cycles.

• Noise: Random variation around the long-term trend.

Regression is one way to extract the trend from a series, as we saw in the previous section. But if the trend is not a simple function, a good alternative is a moving average. A moving average divides the series into overlapping regions, called windows, and computes the average of the values in each

window.One of the simplest moving averages is the rolling mean, which computes the mean of the values in each window. For example, if the window size is 3, the rolling mean computes the mean of values 0 through 2, 1 through 3, 2 through 4, etc.

pandas provides rolling_mean, which takes a Series and a window size and returns a new Series.



```
>>> series = np.arange(10)
array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])

>>> pandas.rolling_mean(series, 3)
array([ nan,  nan,   1,   2,   3,   4,   5,   6,   7,   8])
```

The first two values are nan; the next value is the mean of the first three elements, 0, 1, and 2. The next value is the mean of 1, 2, and 3. And so on.

Before we can apply rolling_mean to the cannabis data, we have to deal with missing values. There are a few days in the observed interval with no reported transactions for one or more quality categories, and a period in 2013 when data collection was not active.

In the DataFrames we have used so far, these dates are absent; the index skips days with no data. For the analysis that follows, we need to represent this missing data explicitly. We can do that by "reindexing" the DataFrame:

```
    dates = pandas.date_range(daily.index.min(), daily.index.max())
    reindexed = daily.reindex(dates)
```

The first line computes a date range that includes every day from the beginning to the end of the observed interval. The second line creates a new DataFrame with all of the data from daily, but including rows for all dates, filled with nan.

Now we can plot the rolling mean like this:

```
    roll_mean = pandas.rolling_mean(reindexed.ppg, 30)
    thinkplot.Plot(roll_mean.index, roll_mean)
```

The window size is 30, so each value in roll_mean is the mean of 30 values from reindexed.ppg.

Figure 12.3 (left) shows the result. The rolling mean seems to do a good job of smoothing out the noise and extracting the trend. The first 29 values are nan, and wherever there's a missing value, it's followed by another 29 nans. There are ways to fill in these gaps, but they are a minor nuisance.

An alternative is the **exponentially-weighted moving average** (EWMA), which has two advantages. First, as the name suggests, it computes a weighted average where the most recent value has the highest weight and the weights for previous values drop off exponentially. Second, the pandas implementation of EWMA handles missing values better.

```
    ewma = pandas.ewma(reindexed.ppg, span=30)
    thinkplot.Plot(ewma.index, ewma)
```

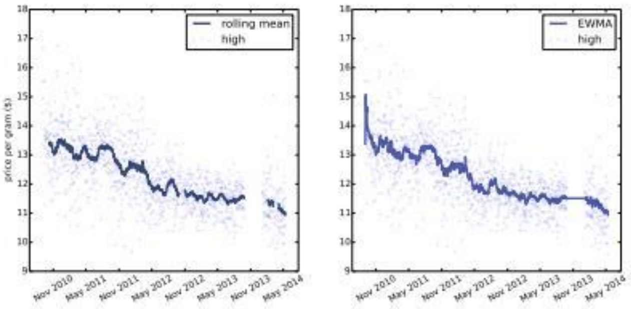


Figure 12.3: Daily price and a rolling mean (left) and exponentially-weighted moving average (right).

The span parameter corresponds roughly to the window size of a moving average; it controls how fast the weights drop off, so it determines the number of points that make a non-negligible contribution to each average.

Figure 12.3 (right) shows the EWMA for the same data. It is similar to the rolling mean, where they are both defined, but it has no missing values, which makes it easier to work with. The values are noisy at the beginning of the time series, because they are based on fewer data points.

Missing values

Time series data based on human behavior often exhibits daily, weekly, monthly, or yearly cycles.

A simple and common way to fill missing data is to use a moving average. The Series method fillna does just what we want:

```
reindexed.ppg.fillna(ewma, inplace=True)
```

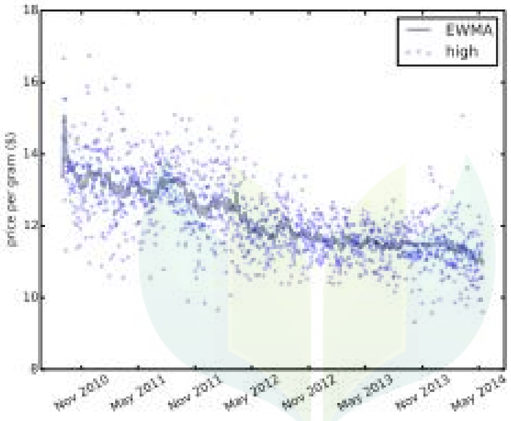
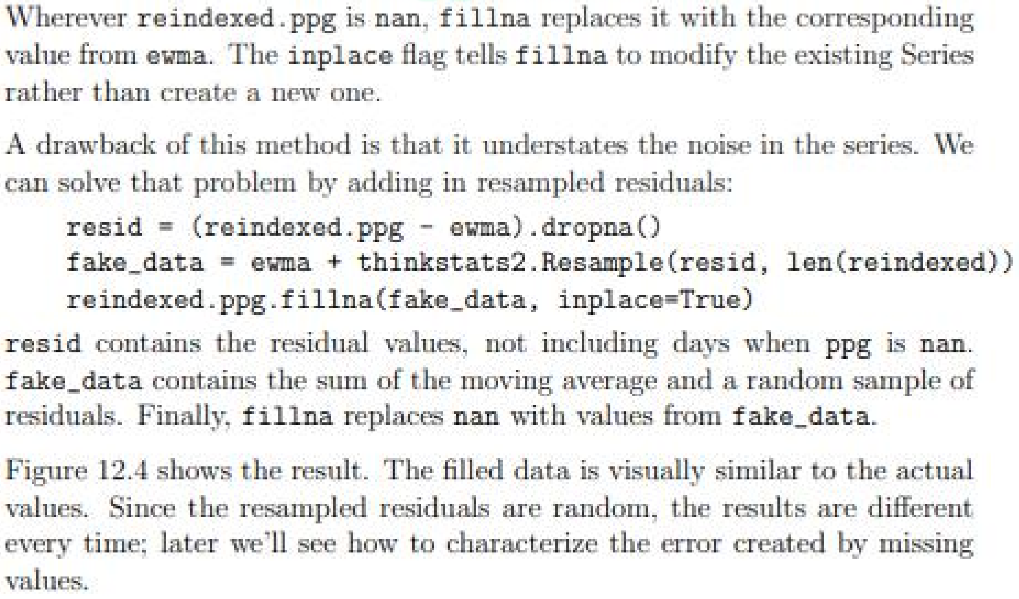


Figure 12.4: Daily price with filled data.



Wherever `reindexed.ppg` is `nan`, `fillna` replaces it with the corresponding value from `ewma`. The `inplace` flag tells `fillna` to modify the existing Series rather than create a new one.

A drawback of this method is that it understates the noise in the series. We can solve that problem by adding in resampled residuals:

```
    resid = (reindexed.ppg - ewma).dropna()
    fake_data = ewma + thinkstats2.Resample(resid, len(reindexed))
    reindexed.ppg.fillna(fake_data, inplace=True)
```

`resid` contains the residual values, not including days when `ppg` is `nan`. `fake_data` contains the sum of the moving average and a random sample of residuals. Finally, `fillna` replaces `nan` with values from `fake_data`.

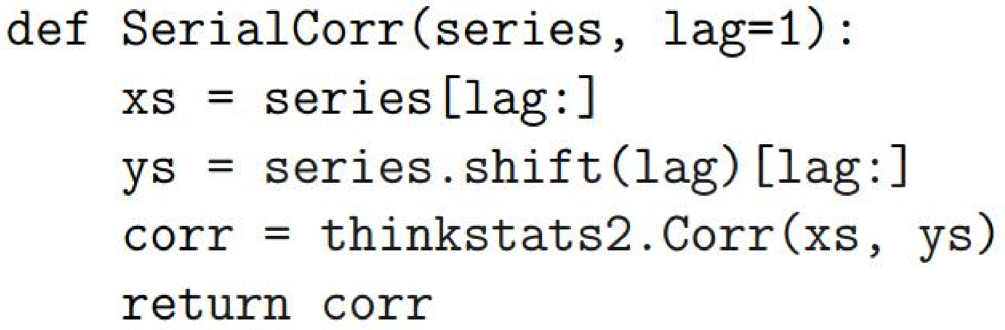
Figure 12.4 shows the result. The filled data is visually similar to the actual values. Since the resampled residuals are random, the results are different every time; later we'll see how to characterize the error created by missing values.

Serial correlation

As prices vary from day to day, you might expect to see patterns. If the price is high on Monday, you might expect it to be high for a few more days; and

Serial correlation

As prices vary from day to day, you might expect to see patterns. If the price is high on Monday, you might expect it to be high for a few more days; and if it's low, you might expect it to stay low. A pattern like this is called serial correlation, because each value is correlated with the next one in the series. To compute serial correlation, we can shift the time series by an interval called a lag, and then compute the correlation of the shifted series with the original:



```
def SerialCorr(series, lag=1):
    xs = series[lag:]
    ys = series.shift(lag)[lag:]
    corr = thinkstats2.Corr(xs, ys)
    return corr
```

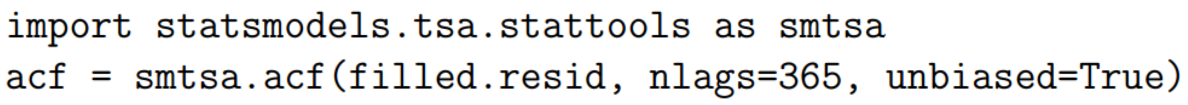
After the shift, the first lag values are nan, so I use a slice to remove them before computing Corr. If we apply SerialCorr to the raw price data with lag 1, we find serial correlation 0.48 for the high quality category, 0.16 for medium and 0.10 for low. In any time series with a long-term trend, we expect to see strong serial correlations; for example, if prices are falling, we expect to see values above the mean in the first half of the series and values below the mean in the second half. It is more interesting to see if the correlation persists if you subtract away the trend. For example, we can compute the residual of the EWMA and then compute its serial correlation:

With lag=1, the serial correlations for the de-trended data are -0.022 for high quality, -0.015 for medium, and 0.036 for low. These values are small, indicating that there is little or no one-day serial correlation in this series.

Autocorrelation

The autocorrelation function is a function that maps from lag to the serial correlation with the given lag. "Autocorrelation" is another name for serial correlation, used more often when the lag is not 1.

StatsModels, which we used for linear regression in Section 11.1, also pro vides functions for time series analysis, including acf, which computes the autocorrelation function:



```
import statsmodels.tsa.stattools as smtsa
acf = smtsa.acf(filled.resid, nlags=365, unbiased=True)
```

acf computes serial correlations with lags from 0 through nlags. The unbiased flag tells acf to correct the estimates for the sample size. The result is an array of correlations. If we select daily prices for high quality, and extract correlations for lags 1, 7, 30, and 365, we can confirm that acf and SerialCorr yield approximately the same results:

Survival analysis

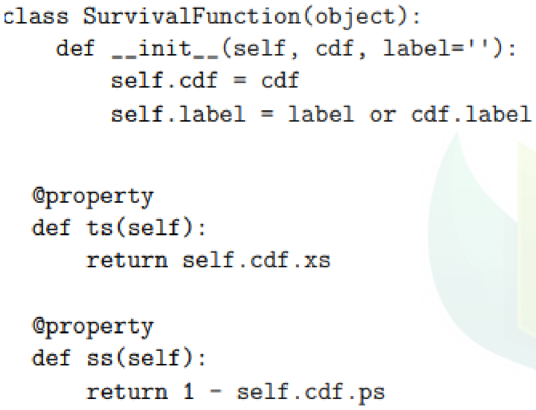
Survival analysis is a way to describe how long things last. It is often used to study human lifetimes, but it also applies to “survival” of mechanical and electronic components, or more generally to intervals in time before an event.

Survival curves

The fundamental concept in survival analysis is the survival curve, S(t), which is a function that maps from a duration, t, to the probability of surviving longer than t. If you know the distribution of durations, or “lifetimes”, finding the survival curve is easy; it’s just the complement of the

CDF: S(t) = 1 − CDF(t)

where CDF(t) is the probability of a lifetime less than or equal to t.



```
class SurvivalFunction(object):
    def __init__(self, cdf, label=''):
        self.cdf = cdf
        self.label = label or cdf.label

  @property
  def ts(self):
      return self.cdf.xs

  @property
  def ss(self):
      return 1 - self.cdf.ps
```

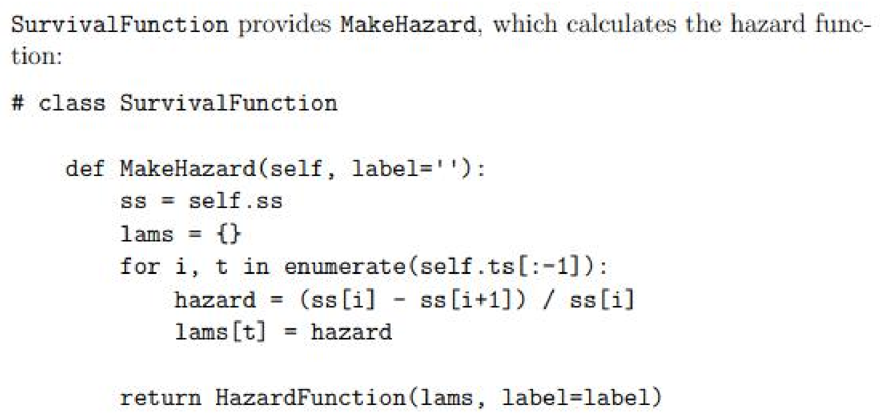
SurvivalFunction provides two properties: ts, which is the sequence of lifetimes, and ss, which is the survival curve.

Hazard function

From the survival curve we can derive the hazard function; for pregnancy lengths, the hazard function maps from a time, t, to the fraction of pregnancies that continue until t and then end at t. To be more precise

$$\lambda(t) = \frac{S(t) - S(t+1)}{S(t)}$$

The numerator is the fraction of lifetimes that end at t, which is also PMF(t).

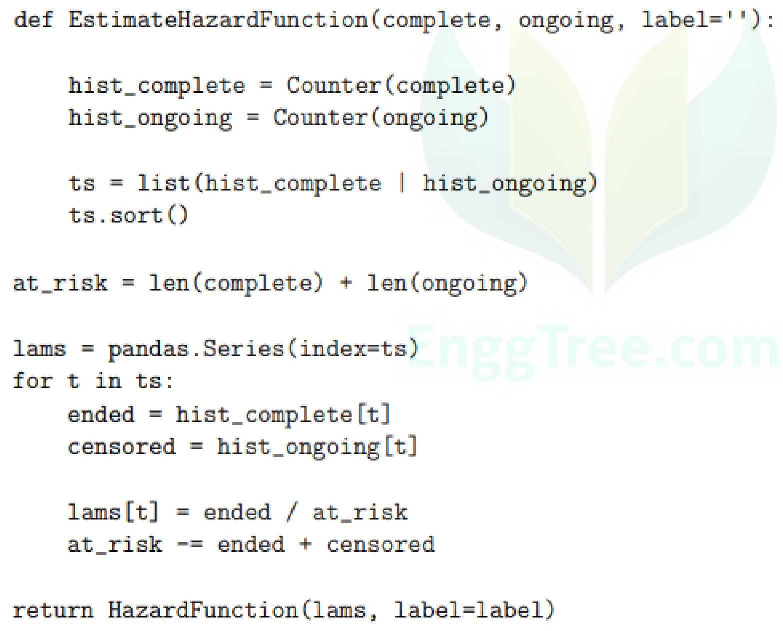


SurvivalFunction provides MakeHazard, which calculates the hazard function:

```
# class SurvivalFunction

    def MakeHazard(self, label=''):
        ss = self.ss
        lams = {}
        for i, t in enumerate(self.ts[:-1]):
            hazard = (ss[i] - ss[i+1]) / ss[i]
            lams[t] = hazard

        return HazardFunction(lams, label=label)
```

Kaplan-Meier estimation

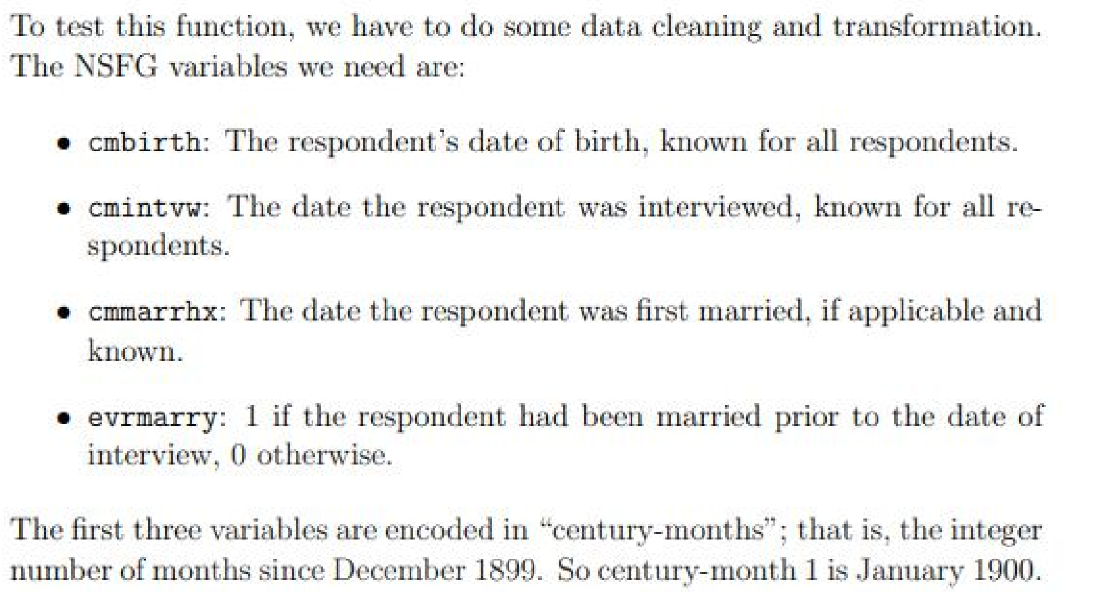
The general idea is that we can use the data to estimate the hazard function, then convert the hazard function to a survival curve. To estimate the hazard function, we consider, for each age, (1) the number of women who got married at that age and (2) the number of women “at risk” of getting married, which includes all women who were not married at an earlier age.



```
def EstimateHazardFunction(complete, ongoing, label=''):

    hist_complete = Counter(complete)
    hist_ongoing = Counter(ongoing)

    ts = list(hist_complete | hist_ongoing)
    ts.sort()

at_risk = len(complete) + len(ongoing)

lams = pandas.Series(index=ts)
for t in ts:
    ended = hist_complete[t]
    censored = hist_ongoing[t]

    lams[t] = ended / at_risk
    at_risk -= ended + censored

return HazardFunction(lams, label=label)
```

complete is the set of complete observations; in this case, the ages when respondents got married. ongoing is the set of incomplete observations; that is, the ages of unmarried women when they were interviewed. First, we precompute hist_complete, which is a Counter that maps from each age to the number of women married at that age, and hist_ongoing which maps from each age to the number of unmarried women interviewed at that age. ts is the union of ages when respondents got married and ages when unmarried women were interviewed, sorted in increasing order. at_risk keeps track of the number of respondents considered “at risk” at each age; initially, it is the total number of respondents.

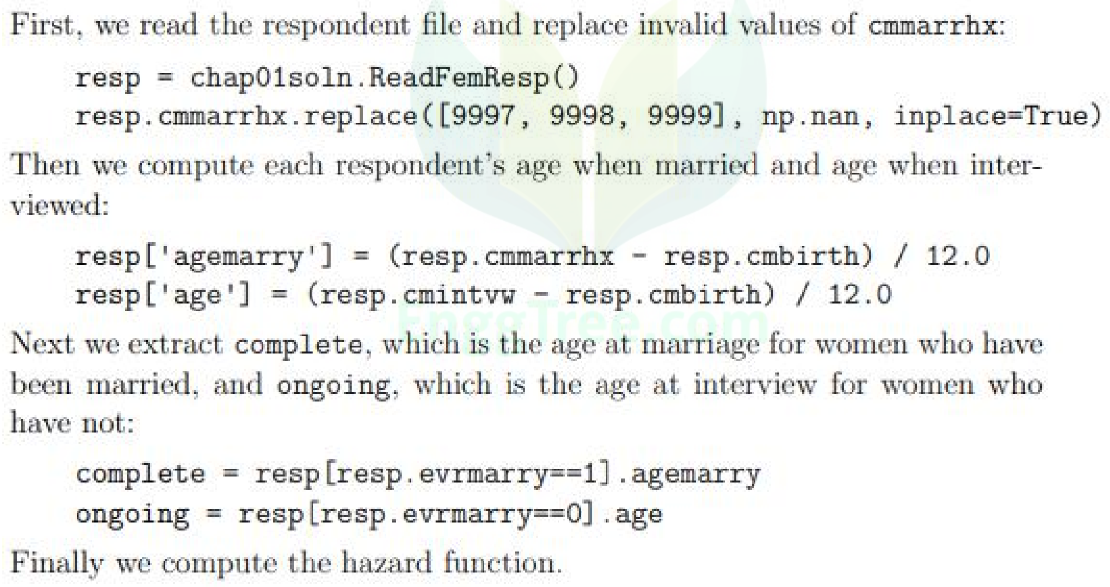
The marriage curve



To test this function, we have to do some data cleaning and transformation. The NSFG variables we need are:

- `cmbirth`: The respondent's date of birth, known for all respondents.
- `cmintvw`: The date the respondent was interviewed, known for all respondents.
- `cmmarrhx`: The date the respondent was first married, if applicable and known.
- `evrmarry`: 1 if the respondent had been married prior to the date of interview, 0 otherwise.

The first three variables are encoded in "century-months"; that is, the integer number of months since December 1899. So century-month 1 is January 1900.



First, we read the respondent file and replace invalid values of `cmmarrhx`:

```
resp = chap01soln.ReadFemResp()
resp.cmmarrhx.replace([9997, 9998, 9999], np.nan, inplace=True)
```

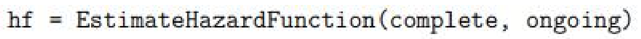
Then we compute each respondent's age when married and age when interviewed:

```
resp['agemarry'] = (resp.cmmarrhx - resp.cmbirth) / 12.0
resp['age'] = (resp.cmintvw - resp.cmbirth) / 12.0
```

Next we extract `complete`, which is the age at marriage for women who have been married, and `ongoing`, which is the age at interview for women who have not:

```
complete = resp[resp.evrmarry==1].agemarry
ongoing = resp[resp.evrmarry==0].age
```

Finally we compute the hazard function.



```
hf = EstimateHazardFunction(complete, ongoing)
```

Cohort effects

Groups like this, defined by date of birth or similar events, are called cohorts, and differences between the groups are called cohort effects.