SASURIE
INSTITUTIONS
COIMBATORE | TIRUPPUR | VIJAYAMANGALAM

SASURIE COLLEGE OF ENGINEERING

DEPARTMENT OF
ARTIFICIAL INTELLIGENCE
AND
DATA SCIENCE

SUBJECT: Fundamentals of Data Science and Analytics

AD3491 Fundamentals of Data Science and Analytics

UNIT I

1) Introduction to Data Science

a) What is Data Science?

Data Science is a combination of multiple disciplines that uses statistics, data analysis, and machine learning to analyze data and to extract knowledge and insights from it.

Data Science is about data gathering, analysis and decision-making.

Data Science is about finding patterns in data, through analysis, and make future predictions.

By using Data Science, companies are able to make:

- Better decisions (should we choose A or B)
- Predictive analysis (what will happen next?)
- Pattern discoveries (find pattern, or maybe hidden information in the data)

b) Where is Data Science Needed?

Data Science is used in many industries in the world today, e.g. banking, consultancy, healthcare, and manufacturing.

Examples of where Data Science is needed:

- For route planning: To discover the best routes to ship
- To foresee delays for flight/ship/train etc. (through predictive analysis)
- To create promotional offers
- To find the best suited time to deliver goods
- To forecast the next years revenue for a company
- To analyze health benefit of training
- To predict who will win elections

Data Science can be applied in nearly every part of a business where data is available. Examples are:

- Consumer goods
- Stock markets
- Industry
- Politics
- Logistic companies
- E-commerce

c) What is Data?

Unstructured data is not organized. We must organize the data for analysis purposes.

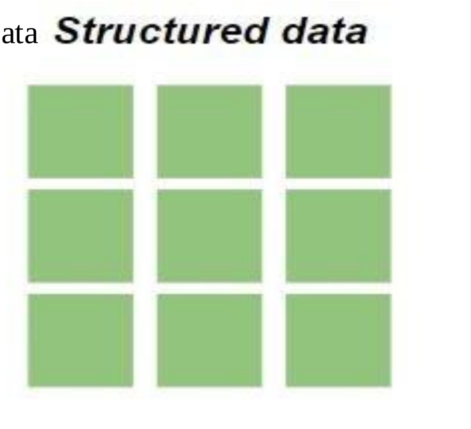
Data is a collection of information.

One purpose of Data Science is to structure data, making it interpretable and easy to work

with.Data can be categorized into two groups:

- Structured data
- Unstructured

dataStructured Data



Unstructured Data

Example of Unstructured data

d) How to Structure Data?

We can use an array or a database table to structure or present data.

Example of an array:

```
[80, 85, 90, 95, 100, 105, 110, 115, 120, 125]
```

The following example shows how to create an array in Python:

```python
Array = [80, 85, 90, 95, 100, 105, 110, 115, 120, 125]
print(Array)
```

e) Data Science & **Python**

Python

Python is a programming language widely used by Data Scientists.

Python has in-built mathematical libraries and functions, making it easier to calculate mathematical problems and to perform data analysis.

We will provide practical examples using Python.

Python Libraries

Python has libraries with large collections of mathematical functions and analytical tools.

In this course, we will use the following libraries:

- Pandas - This library is used for structured data operations, like import CSV files, create dataframes, and data preparation
- Numpy - This is a mathematical library. Has a powerful N-dimensional array object, linear algebra, Fourier transform, etc.
- Matplotlib - This library is used for visualization of data.
- SciPy - This library has linear algebra modules

We will use these libraries throughout the course to create examples.

2) Benefits and uses for Data Science:

1. In Search Engines

The most useful application of Data Science is Search Engines. As we know when we want to search for something on the internet, we mostly used Search engines like Google, Yahoo, Safari, Firefox, etc. So Data Science is used to get Searches faster.

For Example, When we search something suppose "Data Structure and algorithm courses " then at that time on the Internet Explorer we get the first link of GeeksforGeeks Courses. This happens because the GeeksforGeeks website is visited most in order to get information regarding Data Structure courses and Computer related subjects. So this analysis is Done using Data Science, and we get the Topmost visited Web Links.

2. In Transport

Data Science also entered into the Transport field like Driverless Cars. With the help of Driverless Cars, it is easy to reduce the number of Accidents.

For Example, In Driverless Cars the training data is fed into the algorithm and with the help of Data Science techniques, the Data is analyzed like what is the speed limit in Highway, Busy Streets, Narrow Roads, etc. And how to handle different situations while driving etc.

3. In Finance

Data Science plays a key role in Financial Industries. Financial Industries always have an issue of fraud and risk of losses. Thus, Financial Industries needs to automate risk of loss analysis in order to carry out strategic decisions for the company. Also, Financial Industries uses Data Science Analytics tools in order to predict the future. It allows the companies to predict customer lifetime value and their stock market moves.

For Example, In Stock Market, Data Science is the main part. In the Stock Market, Data Science is used to examine past behavior with past data and their goal is to examine the future outcome. Data is analyzed in such a way that it makes it possible to predict future stock prices over a set timetable.

4. In E-Commerce

E-Commerce Websites like Amazon, Flipkart, etc. uses data Science to make a better user experience with personalized recommendations.

For Example, When we search for something on the E-commerce websites we get suggestions similar to choices according to our past data and also we get recommendations according to most buy the product, most rated, most searched, etc. This is all done with the help of Data Science.

5. In Health Care

In the Healthcare Industry data science act as a boon. Data Science is used for:

- Detecting Tumor.
- Drug discoveries.
- Medical Image Analysis.
- Virtual Medical Bots.
- Genetics and Genomics.
- Predictive Modeling for Diagnosis etc.

6. Image Recognition

Currently, Data Science is also used in Image Recognition. **For Example,** When we upload our image with our friend on Facebook, Facebook gives suggestions Tagging who is in the picture. This is done with the help of machine learning and Data Science. When an Image is Recognized, the data analysis is done on one's Facebook friends and after analysis, if the faces which are present in the picture matched with someone else profile then Facebook suggests us auto-tagging.

7. Targeting Recommendation

Targeting Recommendation is the most important application of Data Science. Whatever the user searches on the Internet, he/she will see numerous posts everywhere. This can be explained properly with an example: Suppose I want a mobile phone, so I just Google search it and after that, I changed my mind to buy offline. Data Science helps those companies who are paying for Advertisements for their mobile. So everywhere on the internet in the social media, in the websites, in the apps everywhere I will see the recommendation of that mobile phone which I searched for. So this will force me to buy online.

8. Airline Routing Planning

With the help of Data Science, Airline Sector is also growing like with the help of it, it becomes easy to predict flight delays. It also helps to decide whether to directly land into the destination or take a halt in between like a flight can have a direct route from Delhi to the U.S.A or it can halt in between after that reach at the destination.

9. Data Science in Gaming

In most of the games where a user will play with an opponent i.e. a Computer Opponent, data science concepts are used with machine learning where with the help of past data the Computer will improve its performance. There are many games like Chess, EA Sports, etc. will use Data Science concepts.

10. Medicine and Drug Development

The process of creating medicine is very difficult and time-consuming and has to be done with full disciplined because it is a matter of Someone's life. Without Data Science, it takes lots of time, resources, and finance or developing new Medicine or drug but with the help of Data Science, it becomes easy because the prediction of success rate can be easily determined based on biological data or factors. The algorithms based on data science will forecast how this will react to the human body without lab experiments.

11. In Delivery Logistics

Various Logistics companies like DHL, FedEx, etc. make use of Data Science. Data Science helps these companies to find the best route for the Shipment of their Products, the best time suited for delivery, the best mode of transport to reach the destination, etc.

12. Autocomplete

AutoComplete feature is an important part of Data Science where the user will get the facility to just type a few letters or words, and he will get the feature of auto-completing the line. In Google Mail, when we are writing formal mail to someone so at that time data science concept of Autocomplete feature is used where he/she is an efficient choice to auto-complete the whole line. Also in Search Engines in social media, in various apps, AutoComplete feature is widely used.

3) Facets of Data Science

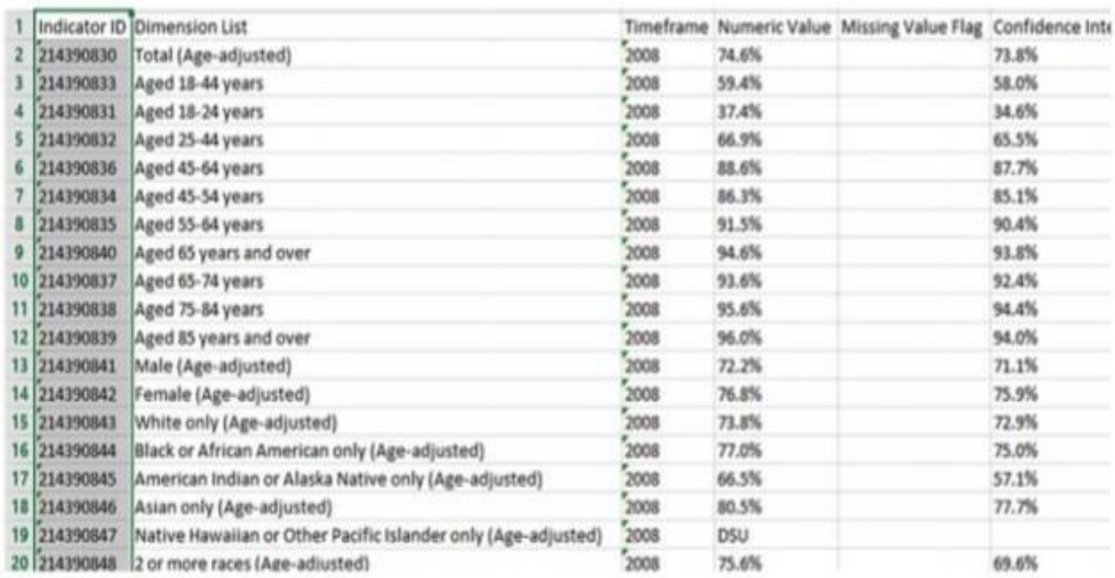
Visualize your data with facets

In Data Science and Big Data you'll come across many different types of data, and each of them tends to require ***different tools and techniques.*** The main categories of data are these:

1. ***Structured***
2. ***Unstructured***
3. ***Natural Language***
4. ***Machine-generated***
5. ***Graph-based***
6. ***Audio, video and images***
7. ***Streaming***

Let’s explore all these interesting data types..

Structured Data



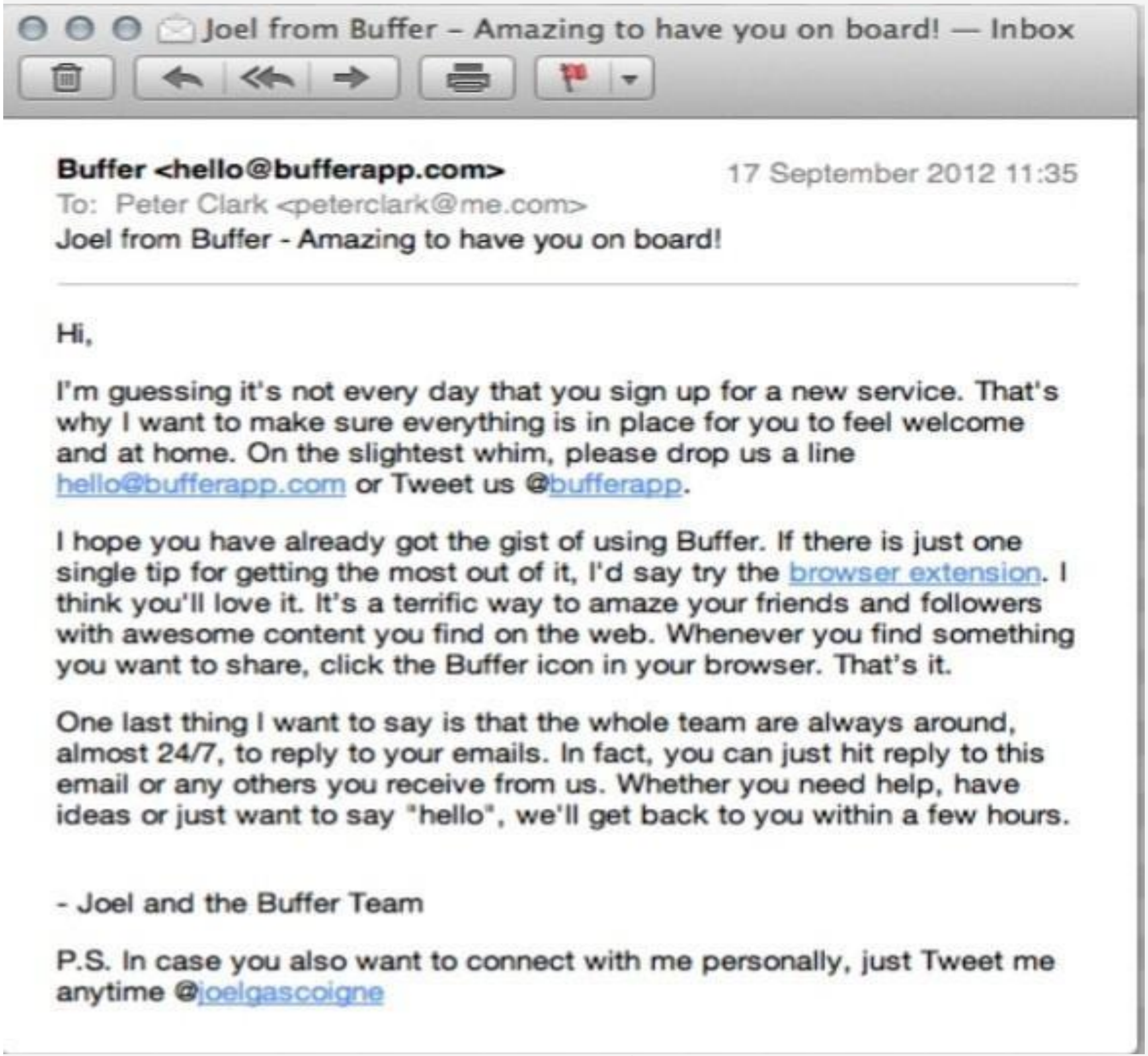
1	Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Inte
2	214390830	Total (Age-adjusted)	2008	74.6%		73.8%
3	214390833	Aged 18-44 years	2008	59.4%		58.0%
4	214390831	Aged 18-24 years	2008	37.4%		34.6%
5	214390832	Aged 25-44 years	2008	66.9%		65.5%
6	214390836	Aged 45-64 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%
19	214390847	Native Hawaiian or Other Pacific Islander only (Age-adjusted)	2008	DSU		
20	214390848	2 or more races (Age-adjusted)	2008	75.6%		69.6%

Fig : An Excel Sheet is an example of Structured Data

Structured data is the data that depends on a data model and resides in a fixed field within a record. It’s often easy to store structured data in ***tables*** within data bases or Excel files. SQL, *Structured Query Language,* is the preferred way to manage and query data that resides in data bases. You may also come across structured data that might give you a hard time storing it in a traditional relational database.

Hierarchical data such as a family tree is one such example.The world isn’t made up of structured data, though; it’s imposed upon it by humans and machines.

Unstructured Data



Buffer <hello@bufferapp.com> 17 September 2012 11:35

To: Peter Clark <peterclark@me.com>

Joel from Buffer - Amazing to have you on board!

Hi,

I'm guessing it's not every day that you sign up for a new service. That's why I want to make sure everything is in place for you to feel welcome and at home. On the slightest whim, please drop us a line hello@bufferapp.com or Tweet us @bufferapp.

I hope you have already got the gist of using Buffer. If there is just one single tip for getting the most out of it, I'd say try the browser extension. I think you'll love it. It's a terrific way to amaze your friends and followers with awesome content you find on the web. Whenever you find something you want to share, click the Buffer icon in your browser. That's it.

One last thing I want to say is that the whole team are always around, almost 24/7, to reply to your emails. In fact, you can just hit reply to this email or any others you receive from us. Whether you need help, have ideas or just want to say "hello", we'll get back to you within a few hours.

- Joel and the Buffer Team

P.S. In case you also want to connect with me personally, just Tweet me anytime @joelgascoigne

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 ***regularemail***. Although email contains structured elements such as the sender, title, and body text, it's a challenge to find the number of people who have written an email complaint about a specific employee because so many ways exist to refer to a person, for example. The thousands of different languages and dialects out there further complicate this. A human-written email, is also a perfect example of 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.***

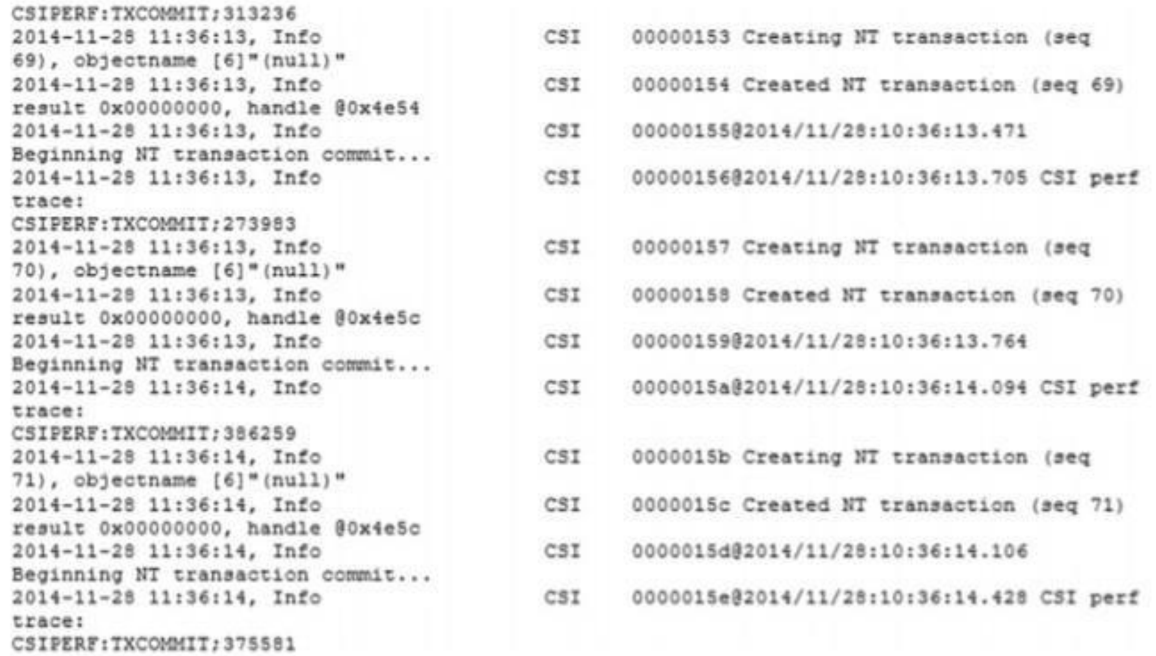
The natural language processing community has had success in ***entity recognition, topic recognition, summarization, text completion, and sentiment analysis***, but models trained in one domain don't generalize well to other domains. Even state-of-the-art techniques aren't able to decipher the meaning of every piece of text. This shouldn't be a surprise though: humans struggle with natural language as well. It's ambiguous by nature. The concept of meaning itself is questionable here. Have two people listen to the same conversation. Will they get the same meaning? The meaning of the same words can vary when coming from someone upset or joyous.

Machine-generated Data

Machine-generated data is informative that's ***automatically created by a computer, process, application or other machine without human intervention.*** Machine-generated data is becoming a major data resource and will continue to do so.

The analysis of Machine data relies on highly scalable tools, due to high volume and speed.

Examples are, *web server logs, call detail records, network event logs and telemetry*.



```
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
```

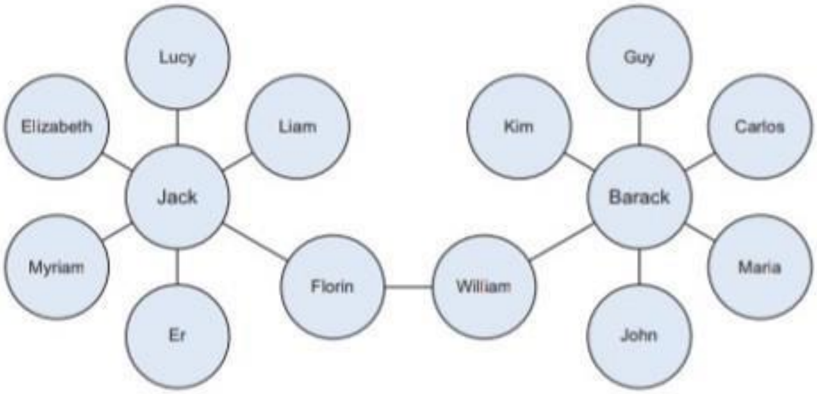
Example for Machine data

This is not the best approach for highly interconnected or “networked” data, where the relationship between entities have a valuable role to play.

Graph-based or Network Data

“Graph data” can be a confusing term because any data can be shown in a graph. “Graph” in this case points to ***mathematical graph theory.*** In graph theory, a graph is a ***mathematical structure to model pair-wise relationships between objects. Graph or network data is, in short, data that focuses on the relationship or adjacency of objects.***

The graph structures use ***nodes, edges, and properties to represent and store graphical data.***



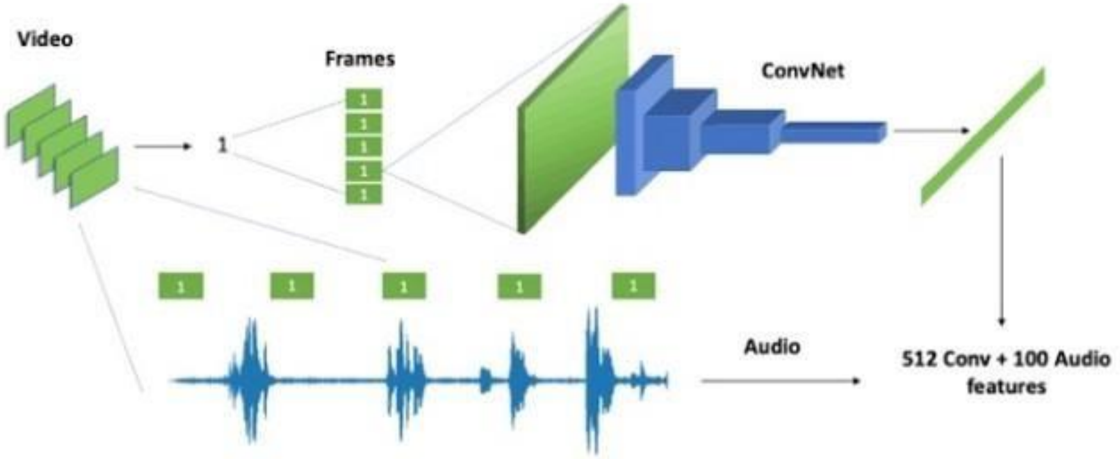
Friends in social network is an example of Graph-based data

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.

Graph databases are used to store graph-based data and are queried with specialized query languages such as **SPARQL.**

Graph data poses its challenges, but for a computer interpreting additive and image data, it can be ever more difficult.

Audio, Images and Videos



Audio, image, and video are data types that pose specific challenges to a data scientist. Tasks that are trivial for humans, such as recognizing objects in pictures, turn out to be challenging for computers.

Multimedia data in the form of audio, video, images and sensor signals have become an integral part of everyday life. Moreover, they have revolutionized product testing and evidence collection by providing multiple sources of data for quantitative and systematic assessment.
We have various libraries, development languages and IDEs commonly used in the field, such as :

- MATLAB
- openCV
- ImageJ
- Python
- R
- Java
- C
- C++
- C#

Streaming Data

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. Although it isn't really a different type of data, we treat it here as much because you need to adapt your process to deal with this type of information.

Examples are the "What's trending" on Twitter, live sporting or music events and the stock market.

4) The data science process:

Overview of the data science process:

The typical data science process consists of six steps:

Figure: 1. The six steps of the data science process:

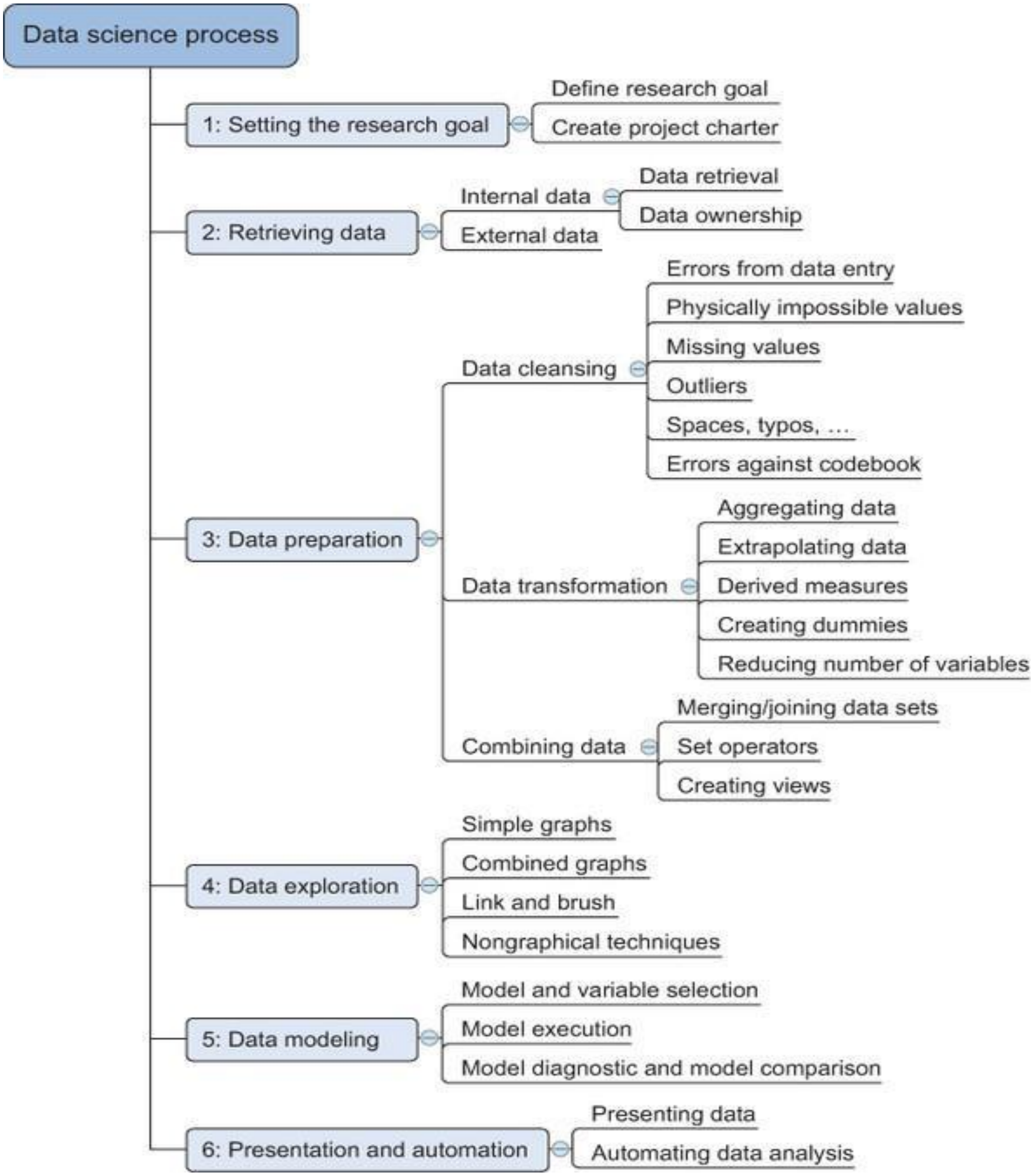


Figure 1 summarizes the data science process and shows the main steps and actions you’ll take during a project.

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. Now that you have the raw data, it’s time to *prepare* it. This 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. Now is the time to bring out the heavy guns, but remember research has taught us that often (but not always) a combination of simple models tends to outperform one complicated model. If you’ve done this phase right, you’re almost done.

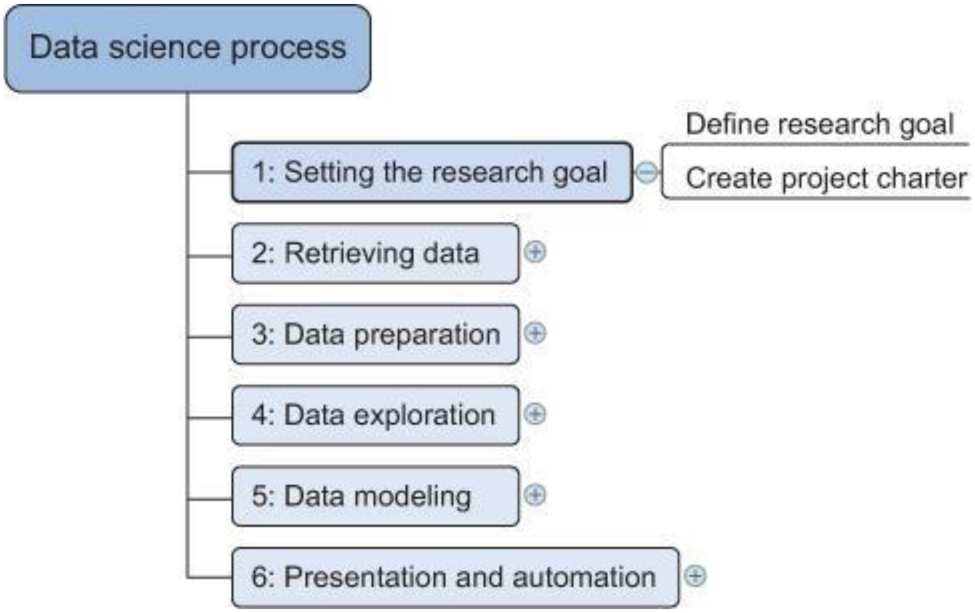
6. The last step of the data science model is *presenting your results and automating the analysis,* if needed. One goal of a project is to change a process and/or make better decisions. You may still need to convince the business that your findings will indeed change the business process as expected. 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.

5)Setting the research goal

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 (figure 2). What does the company expect you to do? And why does management place such a value on your research? Is it part of a bigger strategic picture or a "lone wolf" project originating from an opportunity someone detected? Answering these three questions (what, why, how) is the goal of the first phase, so that everybody knows what to do and can agree on the best course of action.

Figure 2. Step 1: Setting the research goal



The outcome should be a clear research goal, a good understanding of the context, well-defined deliverables, and a plan of action with a timetable. This information is then best placed in a project charter. The length and formality can, of course, differ between projects and companies. In this early phase of the project, people skills and business acumen are more important than great technical prowess, which is why this part will often be guided by more senior personnel.

1. Spend time understanding the goals and context of your research

An essential outcome is the research goal that states the purpose of your assignment in a clear and focused manner. Understanding the business goals and context is critical for project success. 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. Nothing is more frustrating than spending months researching something until you have that one moment of brilliance and solve the problem, but when you report your findings back to the organization, everyone immediately realizes that you misunderstood their question. Don't skim over this phase lightly. Many data scientists fail here: despite their mathematical wit and scientific brilliance, they never seem to grasp the business goals and context.

2. Create a project charter

Clients like to know upfront what they're paying for, so after you have a good understanding of the business problem, try to get a formal agreement on the deliverables. All this information is best collected in a project charter. For any significant project this would be mandatory.

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

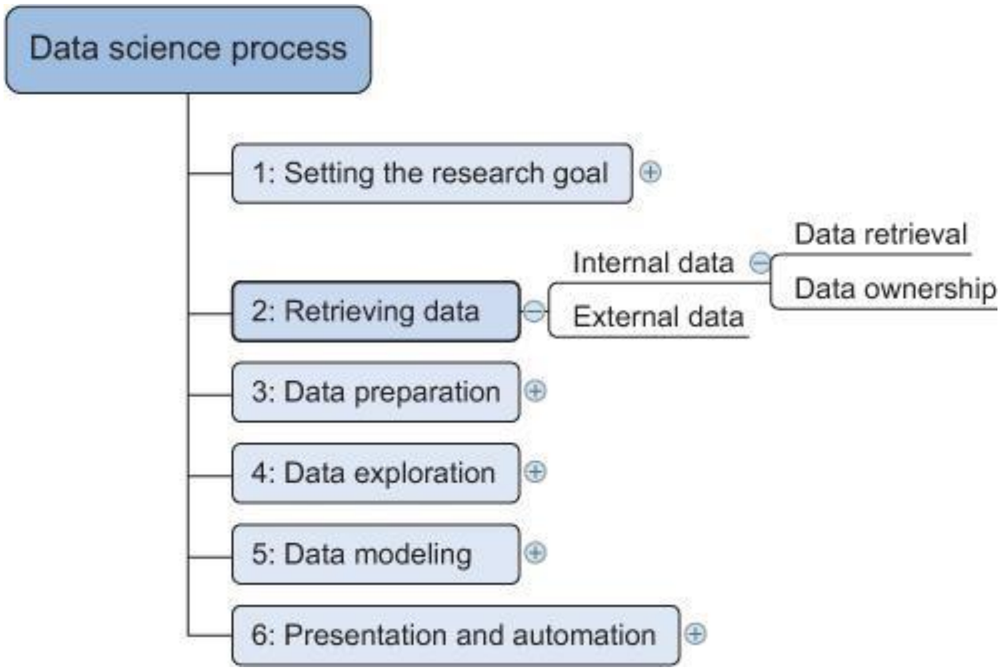
- A timeline

Your client can use this information to make an estimation of the project costs and the data and people required for your project to become a success.

6) Retrieving data

Step 2: Retrieving data

The next step in data science is to retrieve the required data (figure 3). Sometimes you need to go into the field and design a data collection process yourself, but most of the time you won't be involved in this step. Many companies will have already collected and stored the data for you, and what they don't have can often be bought from third parties. Don't be afraid to look outside your organization for data, because more and more organizations are making even high-quality data freely available for public and commercial use.



Data can be stored in many forms, ranging from simple text files to tables in a database. The objective now is acquiring all the data you need. This may be difficult, and even if you succeed, data is often like a diamond in the rough: it needs polishing to be of any use to you.

1. Start with data stored within the company

Your first act should be to assess the relevance and quality of the data that's readily available within your company. Most companies have a program for maintaining key data, so much of the cleaning work may already be done. This data can be stored in official data repositories such as *databases*, *data marts*, *data warehouses*, and *data lakes* maintained by a team of IT professionals. The primary goal of a database is data storage, while a data warehouse is designed for reading and analyzing that data. A data mart is a subset of the data warehouse and geared toward serving a specific business unit. While data warehouses and data marts are home to preprocessed data, data lakes contains data in its natural or raw format. But the possibility exists that your data still resides in Excel files on the desktop of a domain expert.

Finding data even within your own company can sometimes be a challenge. As companies grow, their data becomes scattered around many places. Knowledge of the data may be dispersed as people change positions and leave the company. Documentation and metadata aren't always the top priority of a delivery manager, so it's possible you'll need to develop some Sherlock Holmes–like skills to find all the lost bits.

Getting access to data is another difficult task. Organizations understand the value and sensitivity of data and often have policies in place so everyone has access to what they need and nothing more. These policies translate into physical and digital barriers called *Chinese walls*. These "walls" are mandatory and well-regulated for customer data in most countries. This is for good reasons, too; imagine everybody in a credit card company having access to your spending habits. Getting access to the data may take time and involve company politics.

2. Don't be afraid to shop around

If data isn't available inside your organization, look outside your organization's walls. Many companies specialize in collecting valuable information. For instance, Nielsen and GFK are well known for this in the retail industry. Other companies provide data so that you, in turn, can enrich their services and ecosystem. Such is the case with Twitter, LinkedIn, and Facebook.

Although data is considered an asset more valuable than oil by certain companies, more and more governments and organizations share their data for free with the world. This data can be of excellent quality; it depends on the institution that creates and manages it. The information they share covers a broad range of topics such as the number of accidents or amount of drug abuse in a certain region and its demographics. This data is helpful when you want to enrich proprietary data but also convenient when training your data science skills at home. Table 2.1 shows only a small selection from the growing number of open-data providers.

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

3. Do data quality checks now to prevent problems later

Expect to spend a good portion of your project time doing data correction and cleansing, sometimes up to 80%. The retrieval of data is the first time you’ll inspect the data in the data science process. Most of the errors you’ll encounter during the data-gathering phase are easy to spot, but being too careless will make you spend many hours solving data issues that could have been prevented during data import.

You’ll investigate the data during the import, data preparation, and exploratory phases. The difference is in the goal and the depth of the investigation. During *data retrieval*, you check to see if the data is equal to the data in the source document and look to see if you have the right data types. This shouldn’t take too long; when you have enough evidence that the data is similar to the data you find in the source document, you stop. With *data preparation*, you do a more elaborate check. If you did a good job during the previous phase, the errors you find now are also present in the source document. The focus is on the content of the variables: you want to get rid of typos and other data entry errors and bring the data to a common standard among the data sets. For example, you might correct USQ to USA and United Kingdom to UK. During the *exploratory phase* your focus shifts to what you can learn from the data. Now you assume the data to be clean and look at the statistical properties such as distributions, correlations, and outliers. You’ll often iterate over these phases. For instance, when you discover outliers in the exploratory phase, they can point to a data entry error. Now that you understand how the quality of the data is improved during the process, we’ll look deeper into the data preparation step.

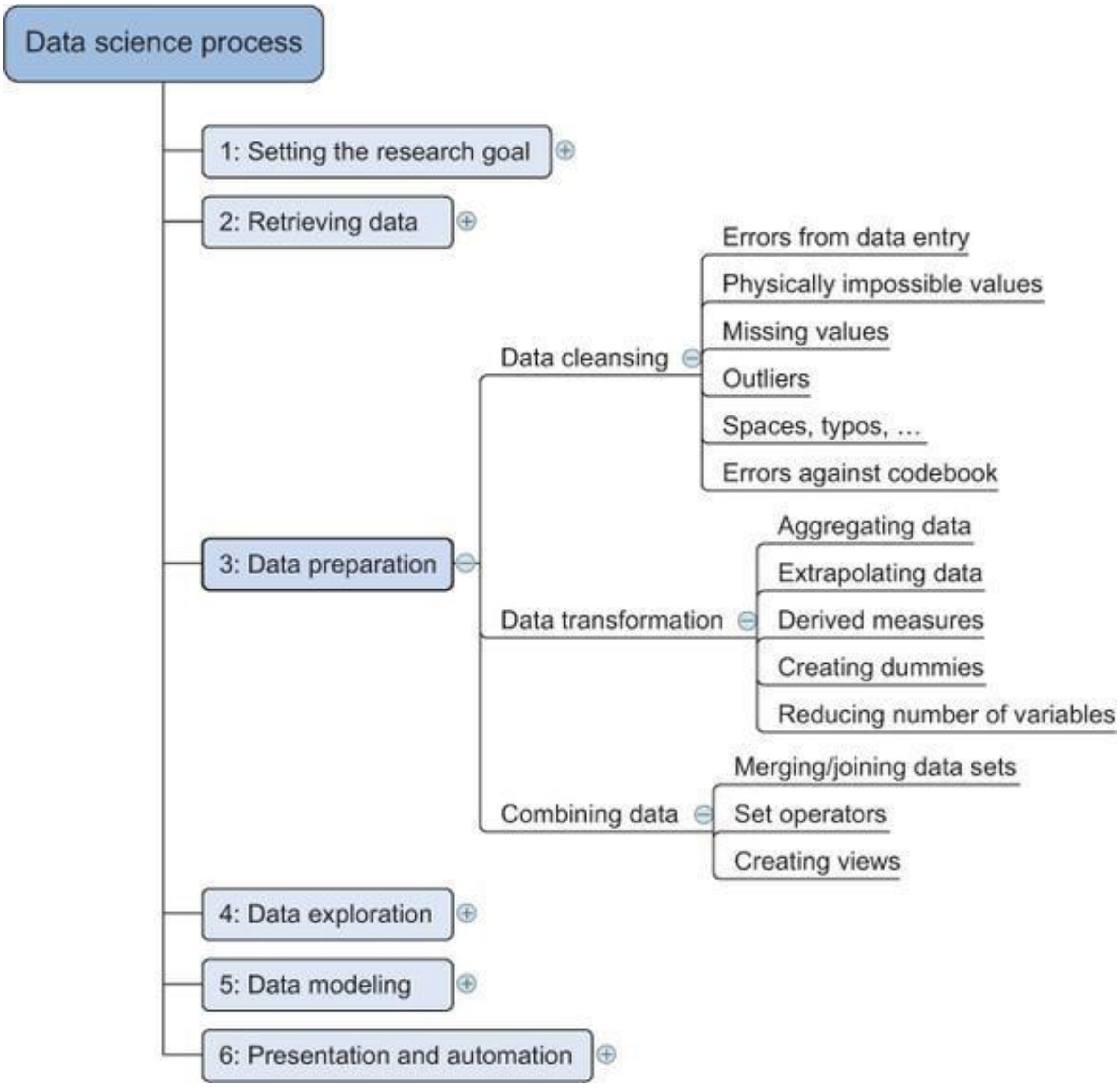
7) Data preparation

Step 3: Cleansing, integrating, and transforming data

The data received from the data retrieval phase is likely to be “a diamond in the rough.” Your task now is to sanitize and prepare it for use in the modeling and reporting phase. Doing so is tremendously important because your models will perform better and you’ll lose less time trying to fix strange output. It can’t be mentioned nearly enough times: garbage in equals garbage out. Your model needs the data in a specific format, so data transformation will always come into play. It’s a good habit to correct data errors as early on in the process as possible. However, this isn’t always possible in a realistic setting, so you’ll need to take corrective actions in your program.

Figure 4 shows the most common actions to take during the data cleansing, integration, and transformation phase.

Figure 4. Step 3: Data preparation:



This mind map may look a bit abstract for now, but we’ll handle all of these points in more detail in the next sections. You’ll see a great commonality among all of these actions.

1. Cleansing data

Data cleansing is a subprocess of the data science process that focuses on removing errors in your data so your data becomes a true and consistent representation of the processes it originates from.

By “true and consistent representation” we imply that at least two types of errors exist. The first type is the *interpretation error*, such as when you take the value in your data for granted, like saying that a person’s age is greater than 300 years. The second type of error points to *inconsistencies* between data sources or against your company’s standardized values. An example of this class of errors is putting “Female” in one table and “F” in another when they represent the same thing: that the person is female. Another example is that you use Pounds in one table and Dollars in another. Too many possible errors exist for this list to be exhaustive, but table 2.2 shows an overview of the types of errors that can be detected with easy checks—the “low hanging fruit,” as it were.

Table 2.2. An overview of common errors

General solution

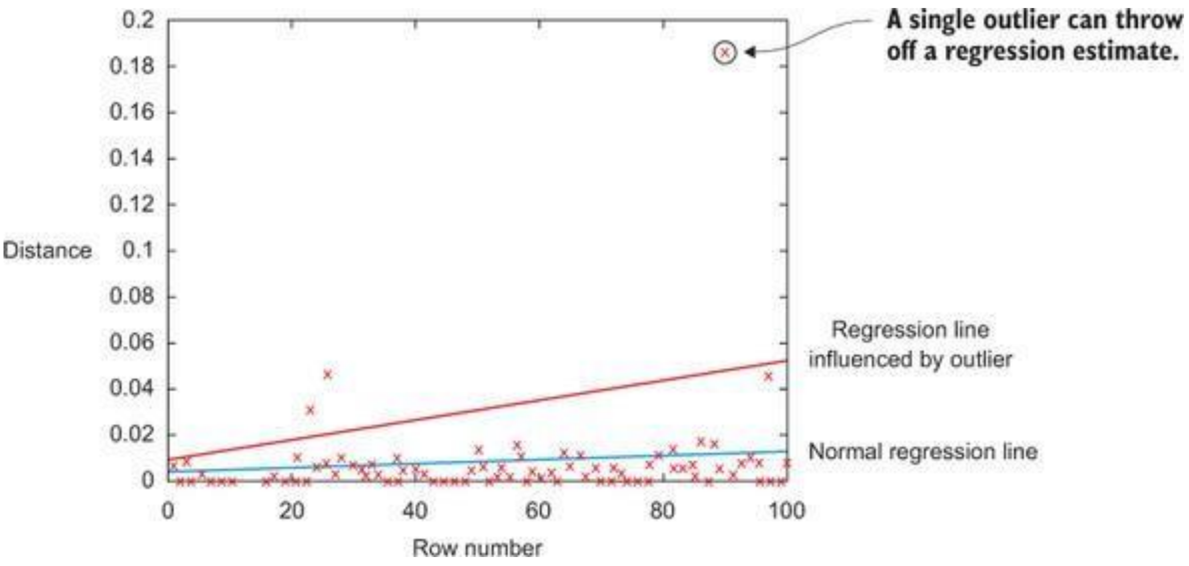
Try to fix the problem early in the data acquisition chain or else fix it in the program.

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

	General solution
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 you'll use more advanced methods, such as simple modeling, to find and identify data errors; diagnostic plots can be especially insightful. For example, in figure 5 we use a measure to identify data points that seem out of place. We do a regression to get acquainted with the data and detect the influence of individual observations on the regression line. When a single observation has too much influence, this can point to an error in the data, but it can also be a valid point. At the data cleansing stage, these advanced methods are, however, rarely applied and often regarded by certain data scientists as overkill.

Figure 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.



Now that we’ve given the overview, it’s time to explain these errors in more detail.

Data entry errors

Data collection and data entry are error-prone processes. They often require human intervention, and because humans are only human, they make typos or lose their concentration for a second and introduce an error into the chain. But data collected by machines or computers isn’t free from errors either. Errors can arise from human sloppiness, whereas others are due to machine or hardware failure. Examples of errors originating from machines are transmission errors or bugs in the extract, transform, and load phase (ETL).

Table 2.3. Detecting outliers on simple variables with a frequency table

Value	Count
Good	1598647
Bad	1354468
Godo	15
Bade	1

Most errors of this type are easy to fix with simple assignment statements and if-then-else rules:

```
if x == "Godo":
    x = "Good"
if x == "Bade":
    x = "Bad"
```

copy

Redundant whitespace

Whitespaces tend to be hard to detect but cause errors like other redundant characters would. Who hasn’t lost a few days in a project because of a bug that was caused by whitespaces at the end of a string? You ask the program to join two keys and notice that observations are missing from the output file. After looking for days through the code, you finally find the bug. Then comes the hardest part: explaining the delay to the project stakeholders. The cleaning during the ETL phase wasn’t well executed, and keys in one table contained a whitespace at the end of a string. This caused a mismatch of keys such as “FR” – “FR”, dropping the observations that couldn’t be matched.

If you know to watch out for them, fixing redundant whitespaces is luckily easy enough in most programming languages. They all provide string functions that will remove the leading and trailing whitespaces. For instance, in Python you can use the `strip()` function to remove leading and trailing spaces.

FIXING CAPITAL LETTER MISMATCHES

Capital letter mismatches are common. Most programming languages make a distinction between “Brazil” and “brazil”. In this case you can solve the problem by applying a function that returns both strings in lowercase, such as `.lower()` in Python. `"Brazil".lower() == "brazil".lower()` should result in `true`.

Impossible values and sanity checks

Sanity checks are another valuable type of data check. Here you check the value against physically or theoretically impossible values such as people taller than 3 meters or someone with an age of 299 years. Sanity checks can be directly expressed with rules:

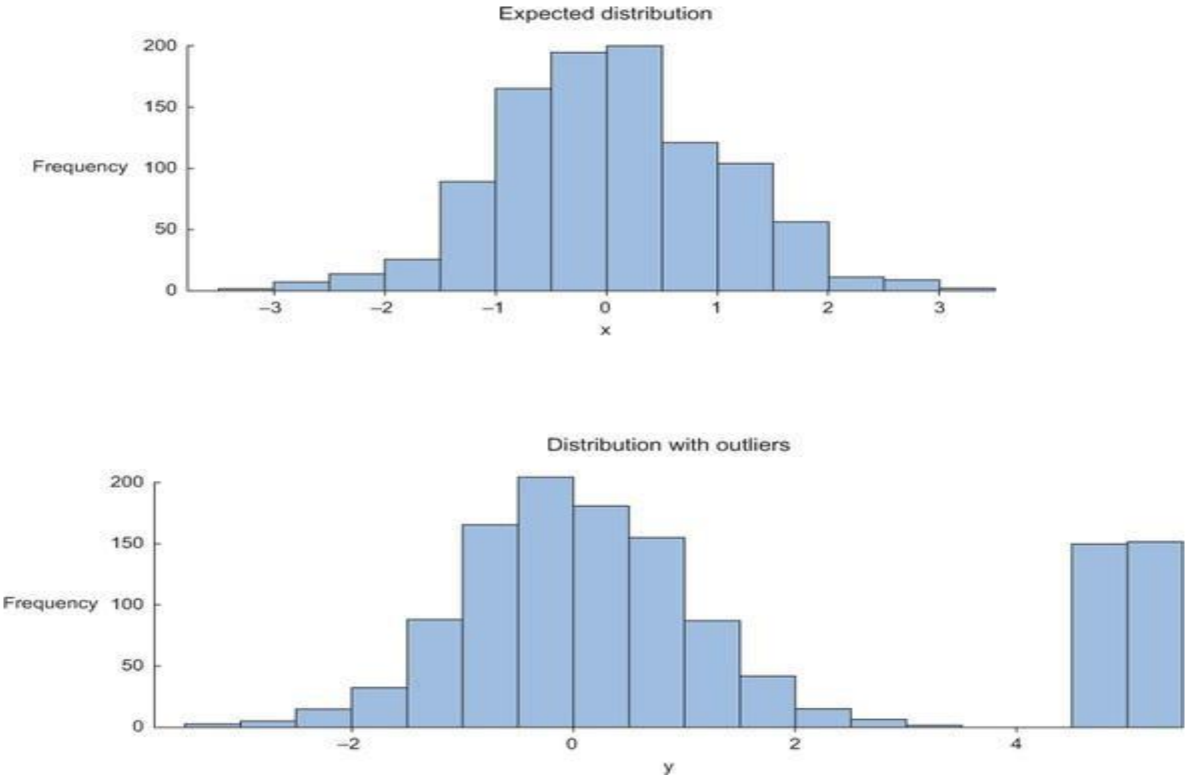
```
check = 0 <= age <= 120
```

copy

Outliers

An outlier is 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. An example is shown in figure 6.

Figure 6. Distribution plots are helpful in detecting outliers and helping you understand the variable.



The plot on the top shows no outliers, whereas the plot on the bottom shows possible outliers on the upper side when a normal distribution is expected. The normal distribution, or Gaussian distribution, is the most common distribution in natural sciences. It shows most cases occurring around the average of the distribution and the occurrences decrease when further away from it. The high values in the bottom graph can point to outliers when assuming a normal distribution. As we saw earlier with the regression example, outliers can gravely influence your data modeling, so investigate them first.

Dealing with missing values

Missing values aren't necessarily wrong, but you still need to handle them separately; certain modeling techniques can't handle missing values. They might be an indicator that something went wrong in your data collection or that an error happened in the ETL process. Common techniques data scientists use are listed in table 2.4.

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	Does not disturb the model as much	Harder to execute You make data assumptions

Technique	Advantage	Disadvantage
distribution		
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

Which technique to use at what time is dependent on your particular case. If, for instance, you don’t have observations to spare, omitting an observation is probably not an option. If the variable can be described by a stable distribution, you could impute based on this. However, maybe a missing value actually means “zero”? This can be the case in sales for instance: if no promotion is applied on a customer basket, that customer’s promo is missing, but most likely it’s also 0, no price cut.

Deviations from a code book

Detecting errors in larger data sets against a code book or against standardized values can be done with the help of set operations. 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. (For instance “0” equals “negative”, “5” stands for “very positive”.) A code book also tells the type of data you’re looking at: is it hierarchical, graph, something else?

You look at those values that are present in set A but not in set B. These are values that should be corrected. It’s no coincidence that *sets* are the data structure that we’ll use when we’re working in code. It’s a good habit to give your data structures additional thought; it can save work and improve the performance of your program.

If you have multiple values to check, it’s better to put them from the code book into a table and use a difference operator to check the discrepancy between both tables. This way, you can profit from the power of a database directly. More on this in chapter 5.

Different units of measurement

When integrating two data sets, you have to pay attention to their respective units of measurement. An example of this would be when you study the prices of gasoline in the world. To do this you gather data from different data providers. Data sets can contain prices per gallon and others can contain prices per liter. A simple conversion will do the trick in this case.

Different levels of aggregation

Having different levels of aggregation is similar to having different types of measurement. 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. But before we tackle this topic we’ll take a little detour and stress the importance of cleaning data as early as possible.

2. Correct errors as early as possible

A good practice is to mediate data errors as early as possible in the data collection chain and to fix as little as possible inside your program while fixing the origin of the problem. Retrieving data is a difficult task, and organizations spend millions of dollars on it in the hope of making better decisions. The data collection process is error-prone, and in a big organization it involves many steps and teams.

Data should be cleansed when acquired for many reasons:

- Not everyone spots the data anomalies. Decision-makers may make costly mistakes on information based on incorrect data from applications that fail to correct for the faulty data.
- 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. For instance, both authors worked at a retailer in the past, and they designed a couponing system to attract more people and make a higher profit. During a data science project, we discovered clients who abused the couponing system and earned money while purchasing groceries. The goal of the couponing system was to stimulate cross-selling, not to give products away for free. This flaw cost the company money and nobody in the company was aware of it. In this case the data wasn't technically wrong but came with unexpected results.
- 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. While doing a small project at a bank we discovered that two software applications used different local settings. This caused problems with numbers greater than 1,000. For one app the number 1.000 meant one, and for the other it meant one thousand.

Fixing the data as soon as it's captured is nice in a perfect world. Sadly, a data scientist doesn't always have a say in the data collection and simply telling the IT department to fix certain things may not make it so. If you can't correct the data at the source, you'll need to handle it inside your code. Data manipulation doesn't end with correcting mistakes; you still need to combine your incoming data.

As a final remark: always keep a copy of your original data (if possible). Sometimes you start cleaning data but you'll make mistakes: impute variables in the wrong way, delete outliers that had interesting additional information, or alter data as the result of an initial misinterpretation. If you keep a copy you get to try again. For "flowing data" that's manipulated at the time of arrival, this isn't always possible and you'll have accepted a period of tweaking before you get to use the data you are capturing. One of the more difficult things isn’t the data cleansing of individual data sets however, it’s combining different sources into a whole that makes more sense.

3. Combining data from different data sources

Your data comes from several different places, and in this substep we focus on integrating these different sources. Data varies in size, type, and structure, ranging from databases and Excel files to text documents.

We focus on data in table structures in this chapter for the sake of brevity. It’s easy to fill entire books on this topic alone, and we choose to focus on the data science process instead of presenting scenarios for every type of data. But keep in mind that other types of data sources exist, such as key-value stores, document stores, and so on, which we’ll handle in more appropriate places in the book.

The different ways of combining data

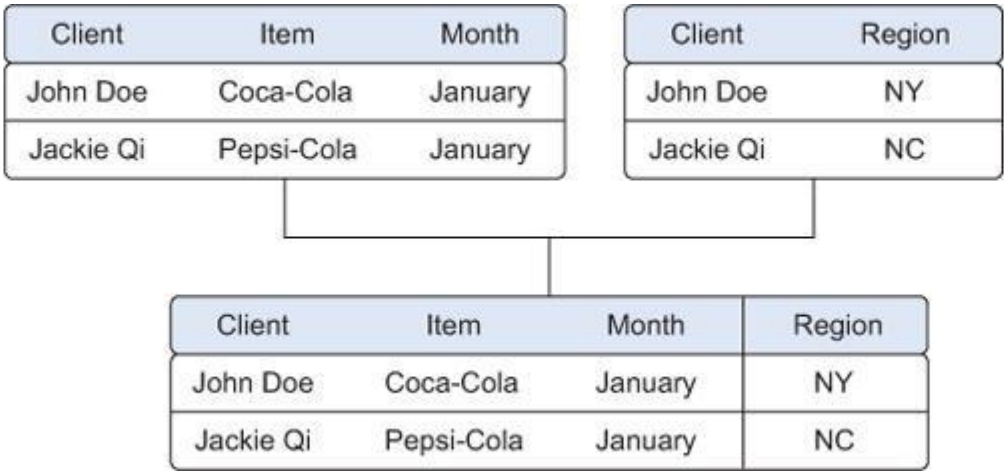
You can perform two operations to combine information from different data sets. The first operation is *joining*: enriching an observation from one table with information from another table. The second operation is *appending* or *stacking*: adding the observations of one table to those of another table.

When you combine data, you have the option to create a new physical table or a virtual table by creating a view. The advantage of a view is that it doesn’t consume more disk space. Let’s elaborate a bit on these methods.

Joining tables

Joining tables allows you to combine the information of one observation found in one table with the information that you find in another table. The focus is on enriching a single observation. Let’s say that the first table contains information about the purchases of a customer and the other table contains information about the region where your customer lives. Joining the tables allows you to combine the information so that you can use it for your model, as shown in figure 2.7.

Figure 7. Joining two tables on the Item and Region keys



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

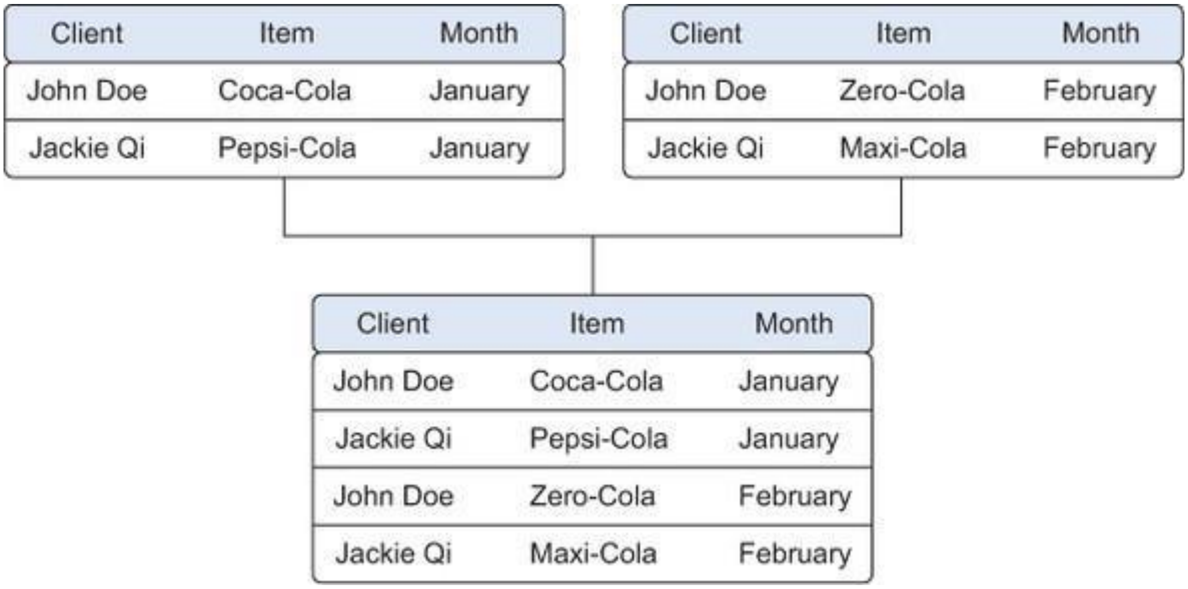
To join tables, you use variables that represent the same object in both tables, such as a date, a country name, or a Social Security number. These common fields are known as keys. When these keys also uniquely define the records in the table they are called *primary keys*. One table may have buying behavior and the other table may have demographic information on a person. In figure 7 both tables contain the client name, and this makes it easy to enrich the client expenditures with the region of the client. People who are acquainted with Excel will notice the similarity with using a lookup function.

The number of resulting rows in the output table depends on the exact join type that you use. We introduce the different types of joins later in the book.

Appending tables

Appending or stacking tables is effectively adding observations from one table to another table. Figure 8 shows an example of appending tables. One table contains the observations from the month January and the second table contains observations from the month February. The result of appending these tables is a larger one with the observations from January as well as February. The equivalent operation in set theory would be the union, and this is also the command in SQL, the common language of relational databases. Other set operators are also used in data science, such as set difference and intersection.

Figure 8. Appending data from tables is a common operation but requires an equal structure in the tables being appended.



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

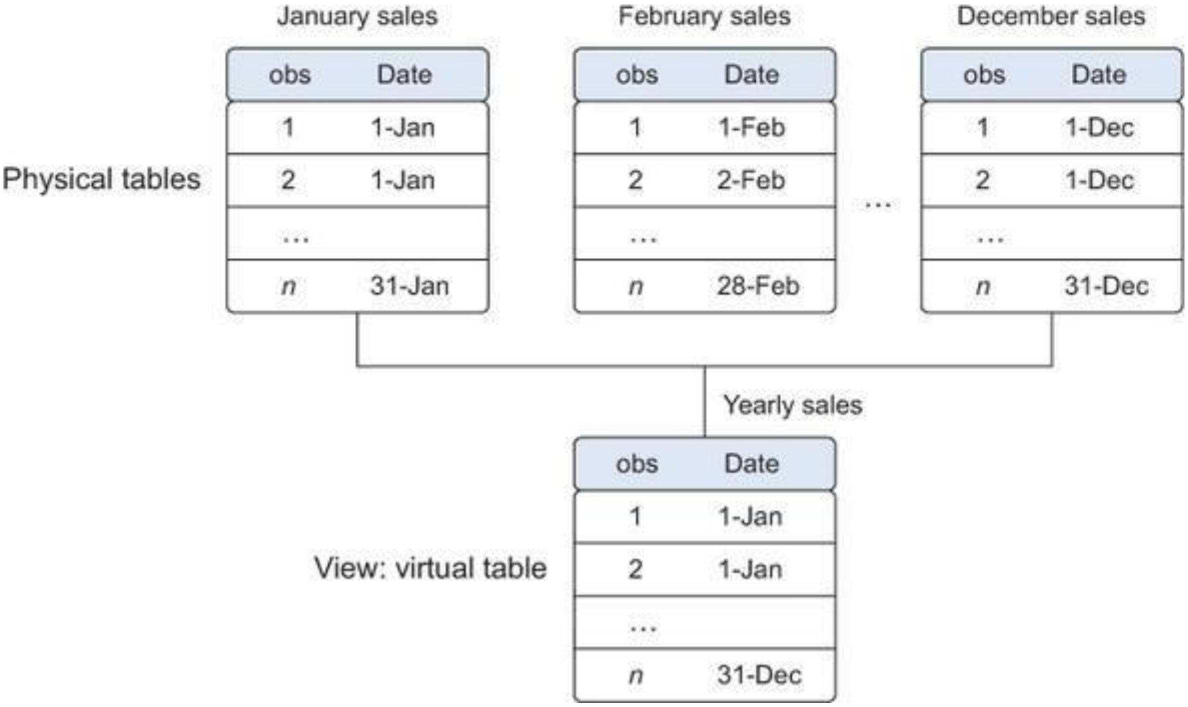
Client	Item	Month
John Doe	Zero-Cola	February
Jackie Qi	Maxi-Cola	February

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

Using views to simulate data joins and appends

To avoid duplication of data, you virtually combine data with views. In the previous example we took the monthly data and combined it in a new physical table. The problem is that we duplicated the data and therefore needed more storage space. In the example we're working with, that may not cause problems, but imagine that every table consists of terabytes of data; then it becomes problematic to duplicate the data. For this reason, the concept of a view was invented. A view behaves as if you're working on a table, but this table is nothing but a virtual layer that combines the tables for you. Figure 9 shows how the sales data from the different months is combined virtually into a yearly sales table instead of duplicating the data. Views do come with a drawback , however. While a table join is only performed once, the join that creates the view is recreated every time it's queried, using more processing power than a pre-calculated table would have.

Figure 9. A view helps you combine data without replication.



Enriching aggregated measures

Data enrichment can also be done by adding calculated information to the table, such as the total number of sales or what percentage of total stock has been sold in a certain region (figure 10)

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

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

Extra measures such as these can add perspective. Looking at figure 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. As always this depends on the exact case, but from our experience models with “relative measures” such as % sales (quantity of product sold/total quantity sold) tend to outperform models that use the raw numbers (quantity sold) as input.

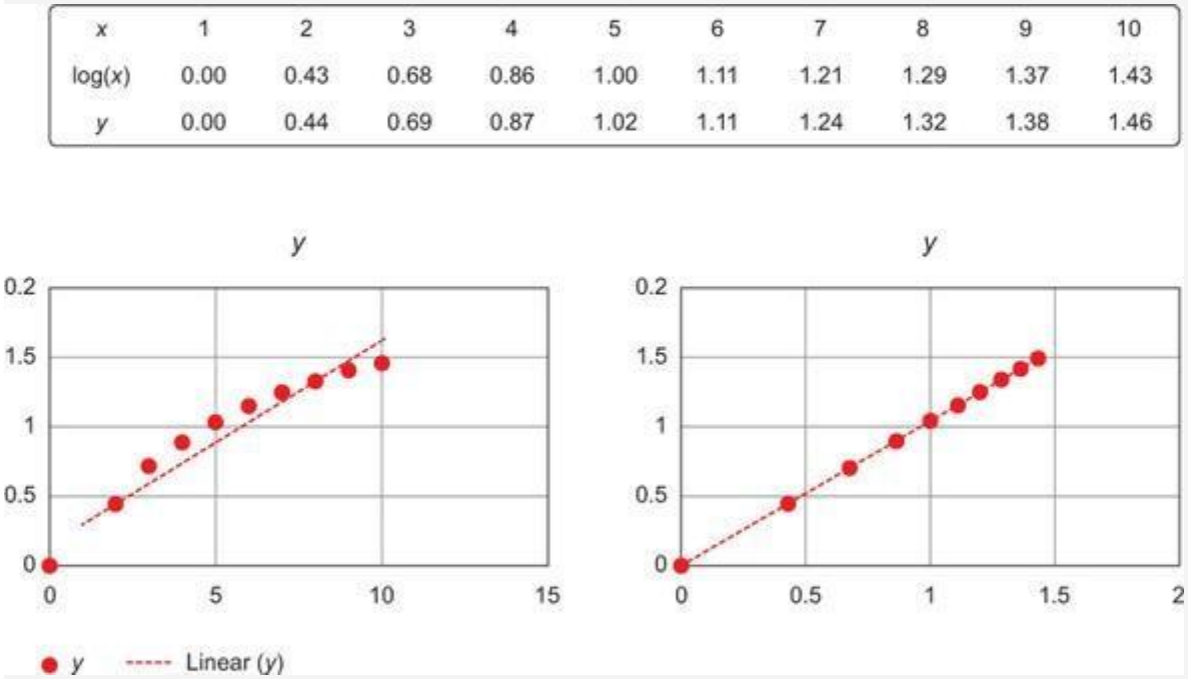
4. Transforming data

Certain models require their data to be in a certain shape. Now that you’ve cleansed and integrated the data, this is the next task you’ll perform: transforming your data so it takes a suitable form for data modeling.

Transforming data

Relationships between an input variable and an output variable aren’t always linear. Take, for instance, a relationship of the form $y = ae^{bx}$. Taking the log of the independent variables simplifies the estimation problem dramatically. Figure 11 shows how transforming the input variables greatly simplifies the estimation problem. Other times you might want to combine two variables into a new variable.

Figure 11. Transforming x to log x makes the relationship between x and y linear (right), compared with the non-log x (left).



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

Reducing the number of variables

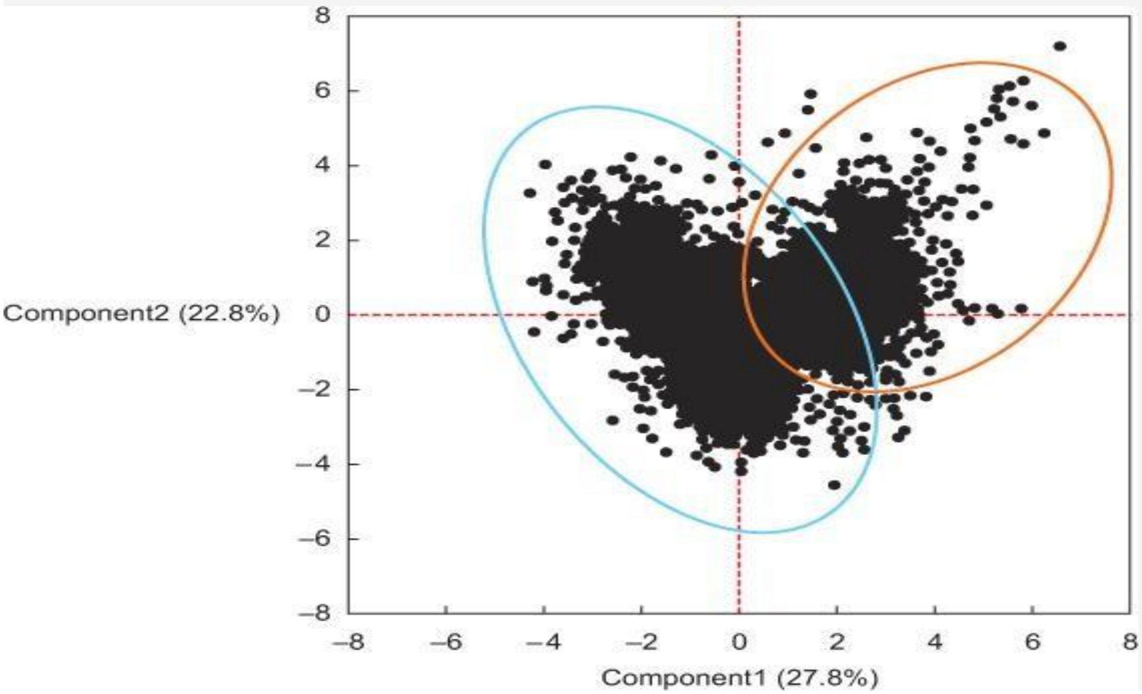
Sometimes you have too many variables and need to reduce the number because they don’t add new information to the model. Having too many variables in your model makes the model difficult to handle, and certain techniques don’t perform well when you overload them with too many input variables. For instance, all the techniques based on a Euclidean distance perform well only up to 10 variables.

EUCLIDEAN DISTANCE

Euclidean distance or “ordinary” distance is an extension to one of the first things anyone learns in mathematics about triangles (trigonometry): Pythagoras’s leg theorem. If you know the length of the two sides next to the 90° angle of a right-angled triangle you can easily derive the length of the remaining side (hypotenuse). The formula for this is hypotenuse = $\sqrt{(\text{side1} + \text{side2})^2}$. The Euclidean distance between two points in a two-dimensional plane is calculated using a similar formula: distance = $\sqrt{((x1 - x2)^2 + (y1 - y2)^2)}$. If you want to expand this distance calculation to more dimensions, add the coordinates of the point within those higher dimensions to the formula. For three dimensions we get distance $= \sqrt{((x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2)}$.

Data scientists use special methods to reduce the number of variables but retain the maximum amount of data. We'll discuss several of these methods in chapter 3. Figure 12 shows how reducing the number of variables makes it easier to understand the key values. It also shows how two variables account for 50.6% of the variation within the data set (component1 = 27.8% + component2 = 22.8%). These variables, called "component1" and "component2," are both combinations of the original variables. They're the *principal components* of the underlying data structure. If it isn't all that clear at this point, don't worry, principal components analysis (PCA) will be explained more thoroughly in chapter 3. What you can also see is the presence of a third (unknown) variable that splits the group of observations into two.

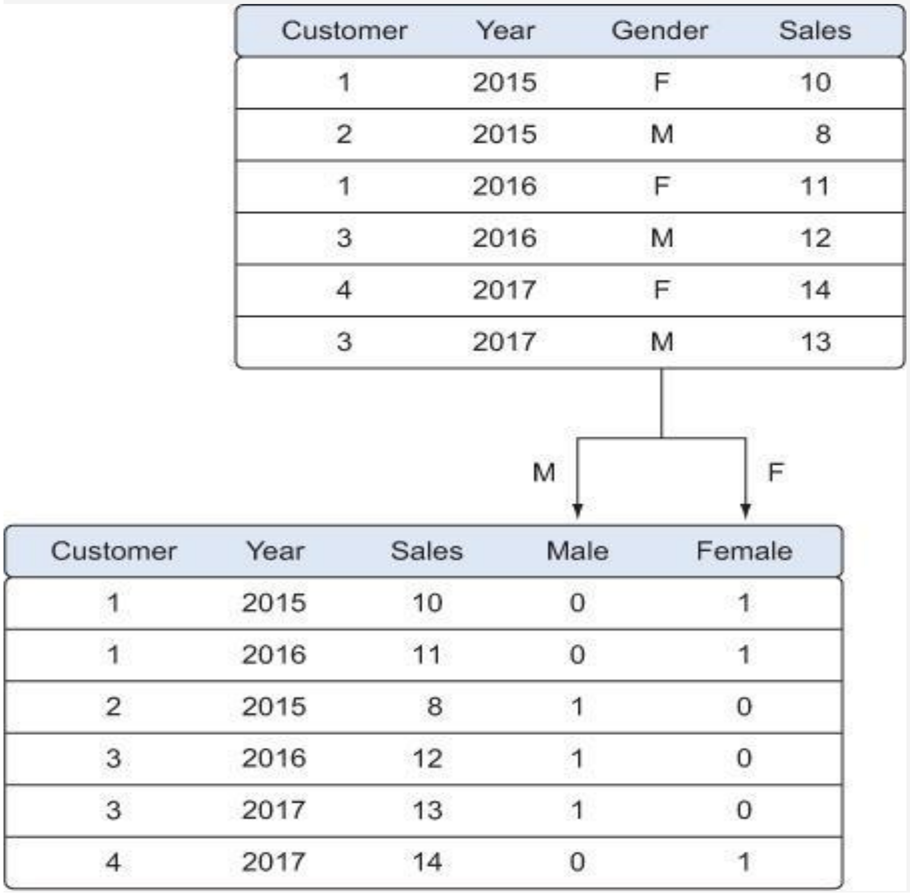
Figure 12. Variable reduction allows you to reduce the number of variables while maintaining as much information as possible.



Turning variables into dummies

Variables can be turned into dummy variables (figure 13). *Dummy variables* can only take two values: true(1) or false(0). They’re used to indicate the absence of a categorical effect that may explain the observation. In this case you’ll make separate columns for the classes stored in one variable and indicate it with 1 if the class is present and 0 otherwise. An example is turning one column named Weekdays into the columns Monday through Sunday. You use an indicator to show if the observation was on a Monday; you put 1 on Monday and 0 elsewhere. Turning variables into dummies is a technique that’s used in modeling and is popular with, but not exclusive to, economists.

Figure 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.



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

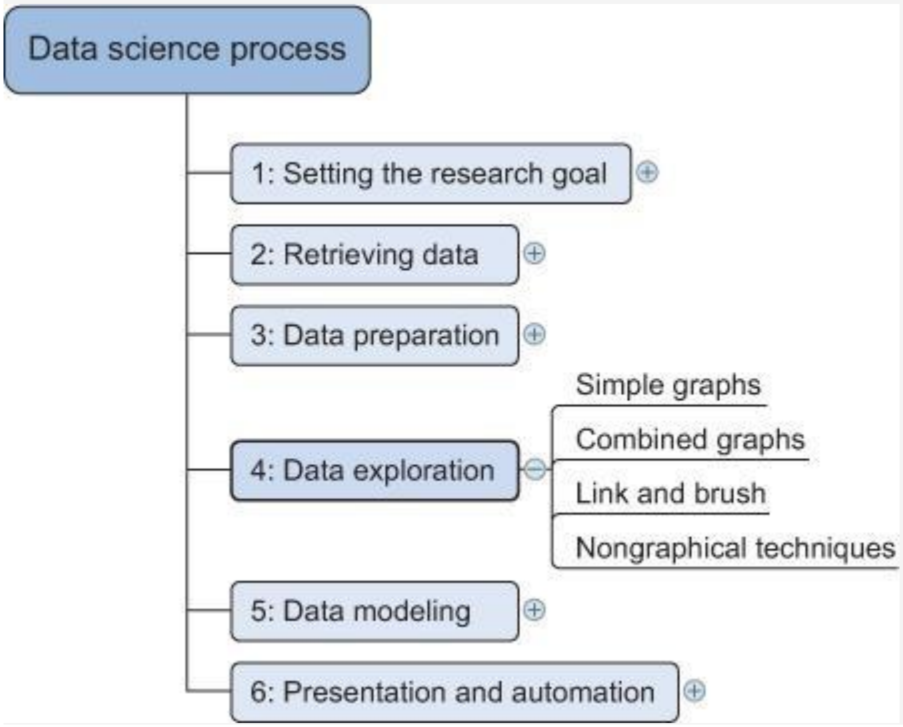
In this section we introduced the third step in the data science process—cleaning, transforming, and integrating data—which changes your raw data into usable input for the modeling phase. The next step in the data science process is to get a better understanding of the content of the data and the relationships between the variables and observations; we explore this in the next section.

8)Data exploration

Step 4: Exploratory data analysis

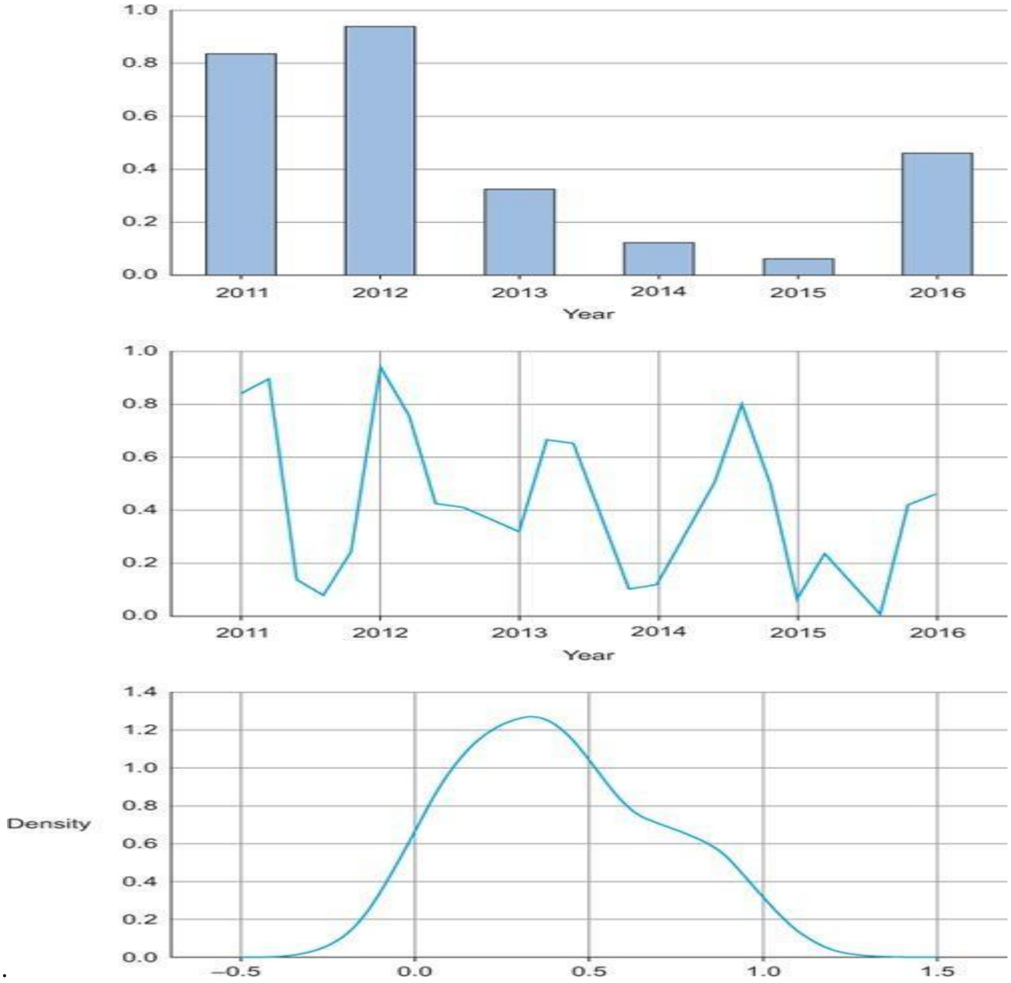
During exploratory data analysis you take a deep dive into the data (see figure 2.14). Information becomes much easier to grasp when shown in a picture, therefore you mainly use graphical techniques to gain an understanding of your data and the interactions between variables. This phase is about exploring data, so keeping your mind open and your eyes peeled is essential during the exploratory data analysis phase. The goal isn't to cleanse the data, but it's common that you'll still discover anomalies you missed before, forcing you to take a step back and fix them.

Figure 14. Step 4: Data exploration



The visualization techniques you use in this phase range from simple line graphs or histograms, as shown in figure 15, to more complex diagrams such as Sankey and network graphs. Sometimes it’s useful to compose a composite graph from simple graphs to get even more insight into the data. Other times the graphs can be animated or made interactive to make it easier and, let’s admit it, way more fun.

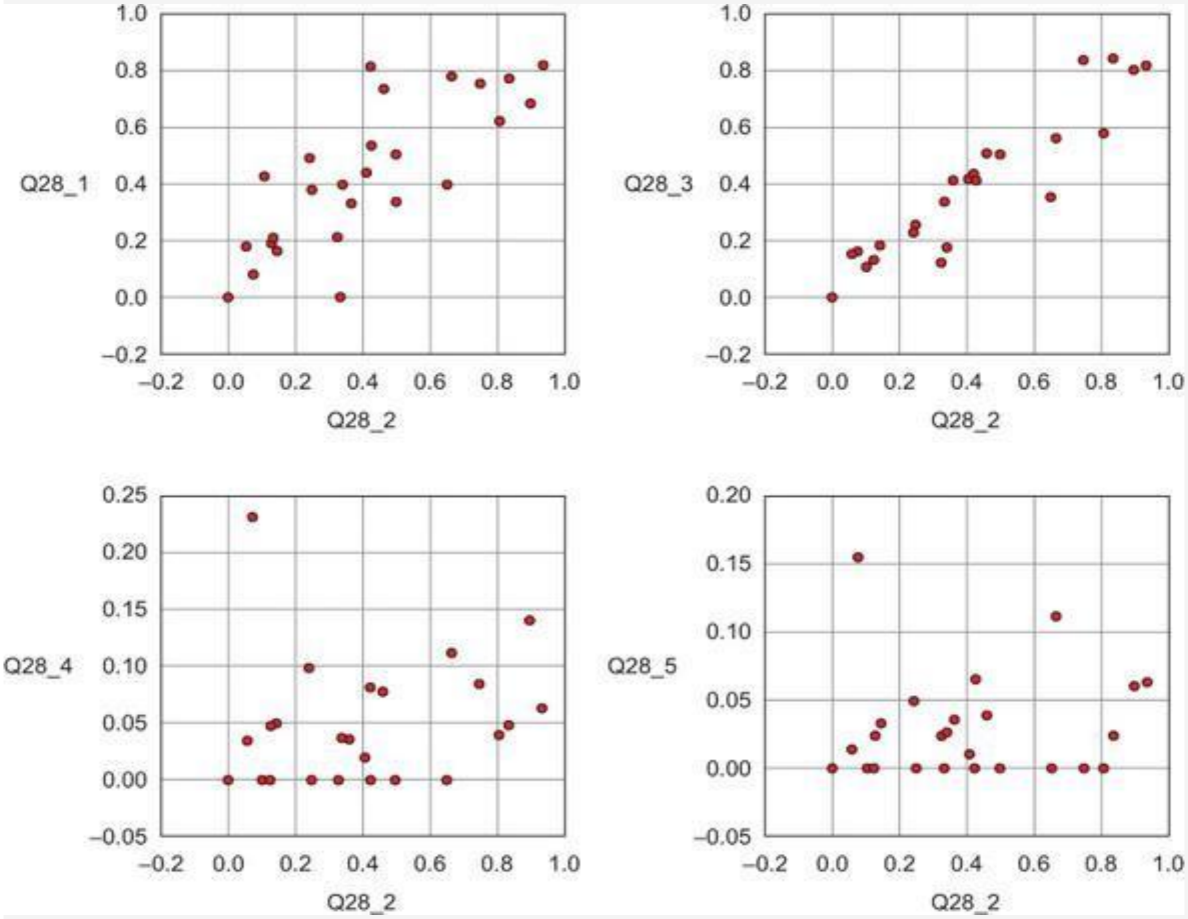
Figure 2.15. From top to bottom, a bar chart, a line plot, and a distribution are some of the graphs used in exploratory analysis.



Mike Bostock has interactive examples of almost any type of graph. It’s worth spending time on his website, though most of his examples are more useful for data presentation than data exploration.

These plots can be combined to provide even more insight, as shown in figure 16.

Figure 16. Drawing multiple plots together can help you understand the structure of your data over multiple variables.



Overlaying several plots is common practice. In figure 17 we combine simple graphs into a Pareto diagram, or 80-20 diagram.

Figure 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%.

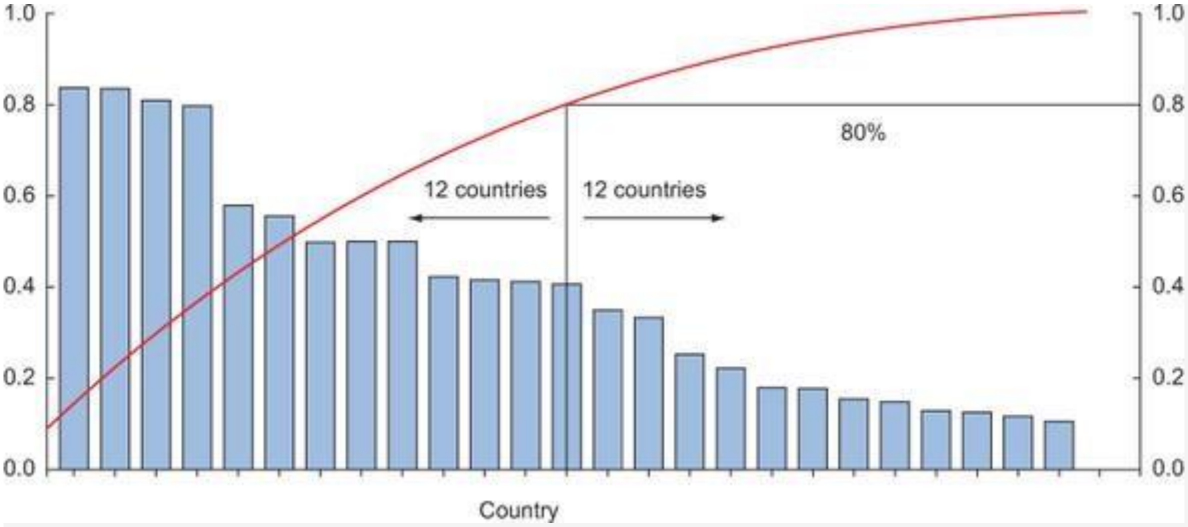


Figure 18 shows another technique: *brushing and linking*. With brushing and linking you combine and link different graphs and tables (or views) so changes in one graph are automatically transferred to the other graphs.

Figure 18. Link and brush allows you to select observations in one plot and highlight the same observations in the other plots.

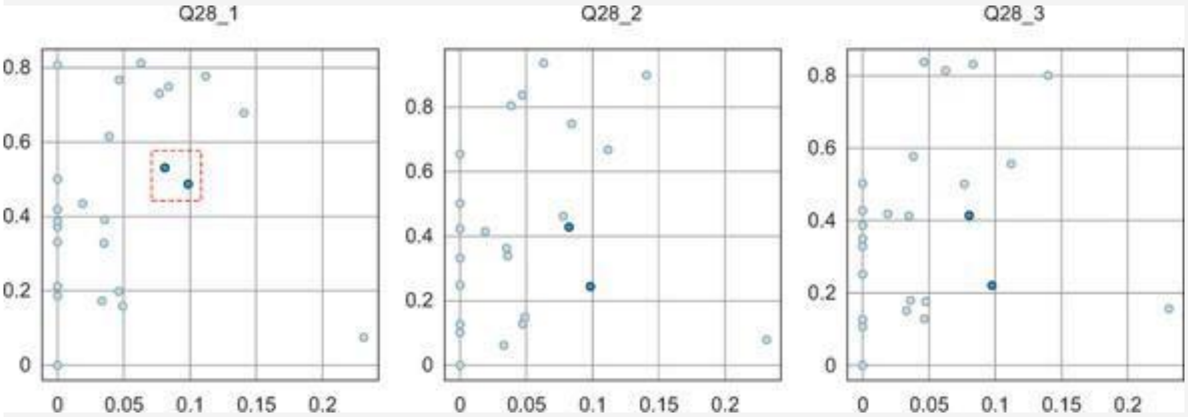


Figure 18 shows the average score per country for questions. Not only does this indicate a high correlation between the answers, but it's easy to see that when you select several points on a subplot, the points will correspond to similar points on the other graphs. In this case the selected points on the left graph correspond to points on the middle and right graphs, although they correspond better in the middle and right graphs.

Two other important graphs are the histogram shown in figure 19 and the boxplot shown in figure 20.

Figure 19. Example histogram: the number of people in the age-groups of 5-year intervals

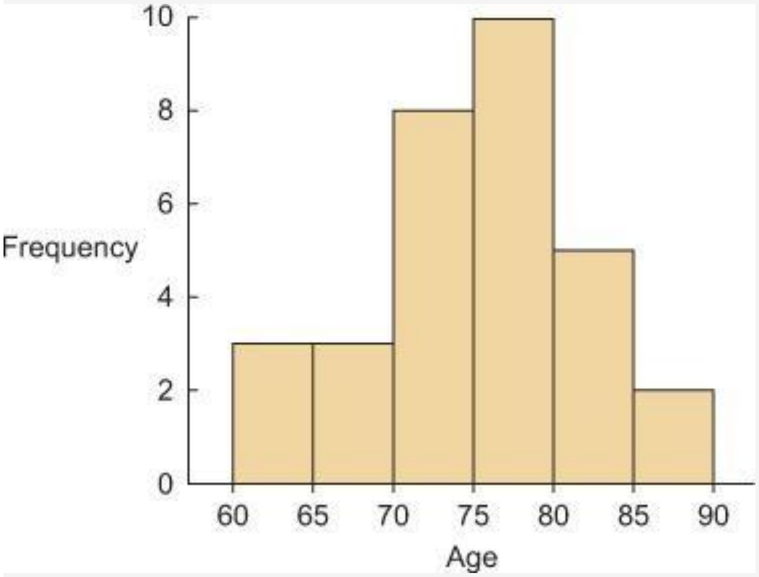
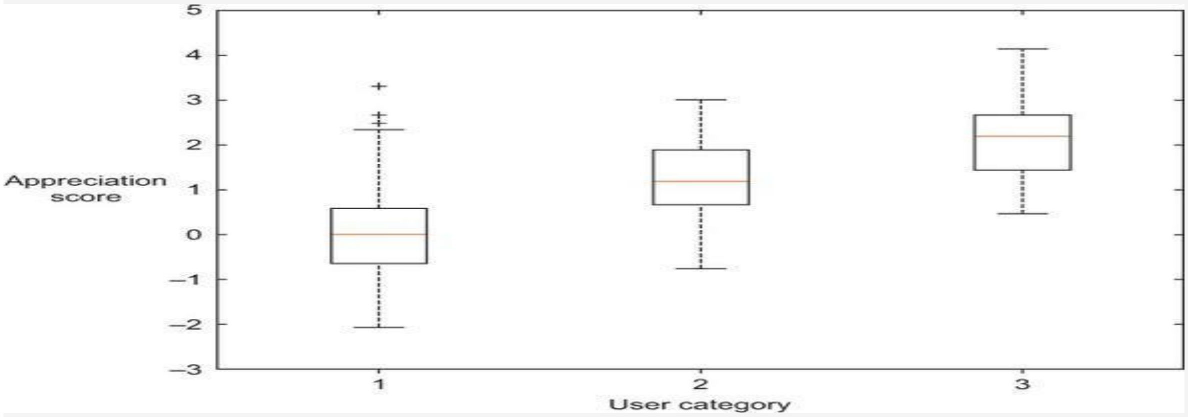


Figure 20. Example boxplot: each user category has a distribution of the appreciation each has for a certain picture on a photography website.



In a histogram a variable is cut into discrete categories and the number of occurrences in each category are summed up and shown in the graph. The boxplot, on the other hand, doesn't show how many observations are present but does offer an impression of the distribution within categories. It can show the maximum, minimum, median, and other characterizing measures at the same time.

The techniques we described in this phase are mainly visual, but in practice they're certainly not limited to visualization techniques. Tabulation, clustering, and other modeling techniques can also be a part of exploratory analysis. Even building simple models can be a part of this step.

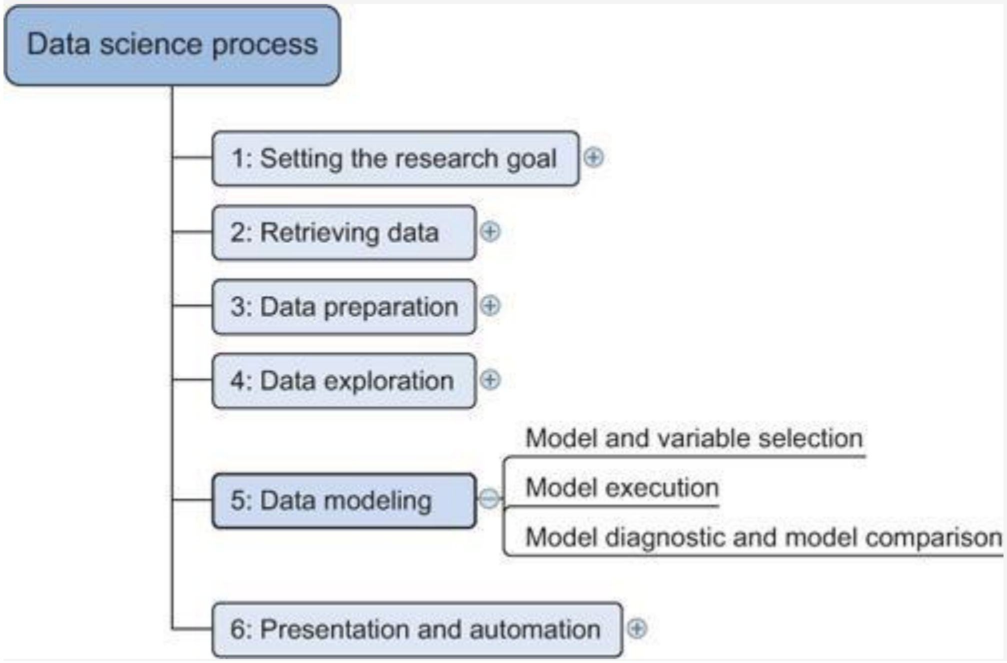
Now that you've finished the data exploration phase and you've gained a good grasp of your data, it's time to move on to the next phase: building models.

9)Data modeling

Step 5: Build the models

With clean data in place and a good understanding of the content, you're ready to build models with the goal of making better predictions, classifying objects, or gaining an understanding of the system that you're modeling. This phase is much more focused than the exploratory analysis step, because you know what you're looking for and what you want the outcome to be. Figure 21 shows the components of model building.

Figure 21. Step 5: Data modeling


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

Building a model is an iterative process. The way you build your model depends on whether you go with classic statistics or the somewhat more recent machine learning school, and the type of technique you want to use. Either way, most models consist of the following main steps:

1. Selection of a modeling technique and variables to enter in the model

2. Execution of the model

3. Diagnosis and model comparison

1. Model and variable selection

You'll need to select the variables you want to include in your model and a modeling technique. Your findings from the exploratory analysis should already give a fair idea of what variables will help you construct a good model. Many modeling techniques are available, and choosing the right model for a problem requires judgment on your part. You'll need to consider model performance and whether your project meets all the requirements to use your model, as well as other factors:

- 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?
- Does the model need to be easy to explain?

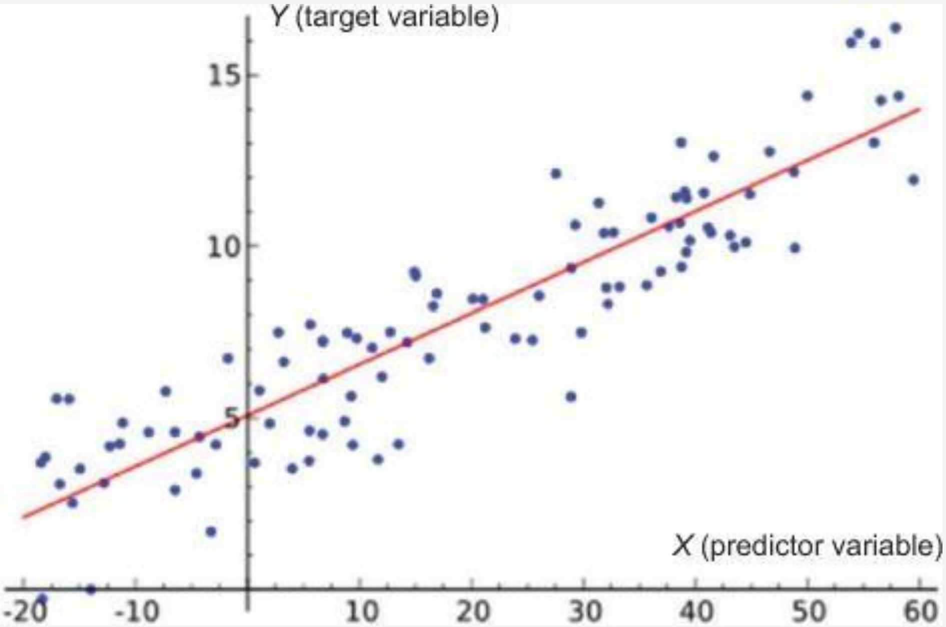
When the thinking is done, it's time for action.

2. Model execution

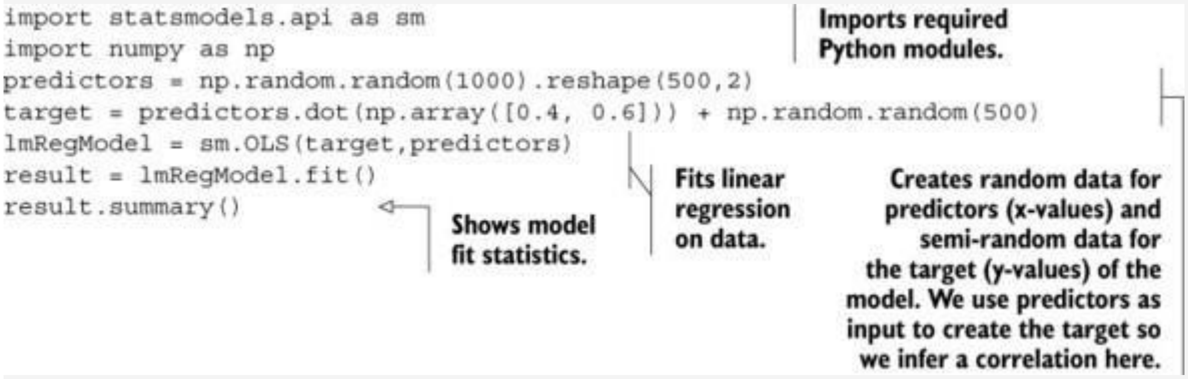
Once you've chosen a model you'll need to implement it in code.

Luckily, most programming languages, such as Python, already have libraries such as StatsModels or Scikit-learn. These packages use several of the most popular techniques. Coding a model is a nontrivial task in most cases, so having these libraries available can speed up the process. As you can see in the following code, it’s fairly easy to use linear regression (figure 22) with StatsModels or Scikit-learn. Doing this yourself would require much more effort even for the simple techniques. The following listing shows the execution of a linear prediction model.

Figure 22. Linear regression tries to fit a line while minimizing the distance to each point



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.

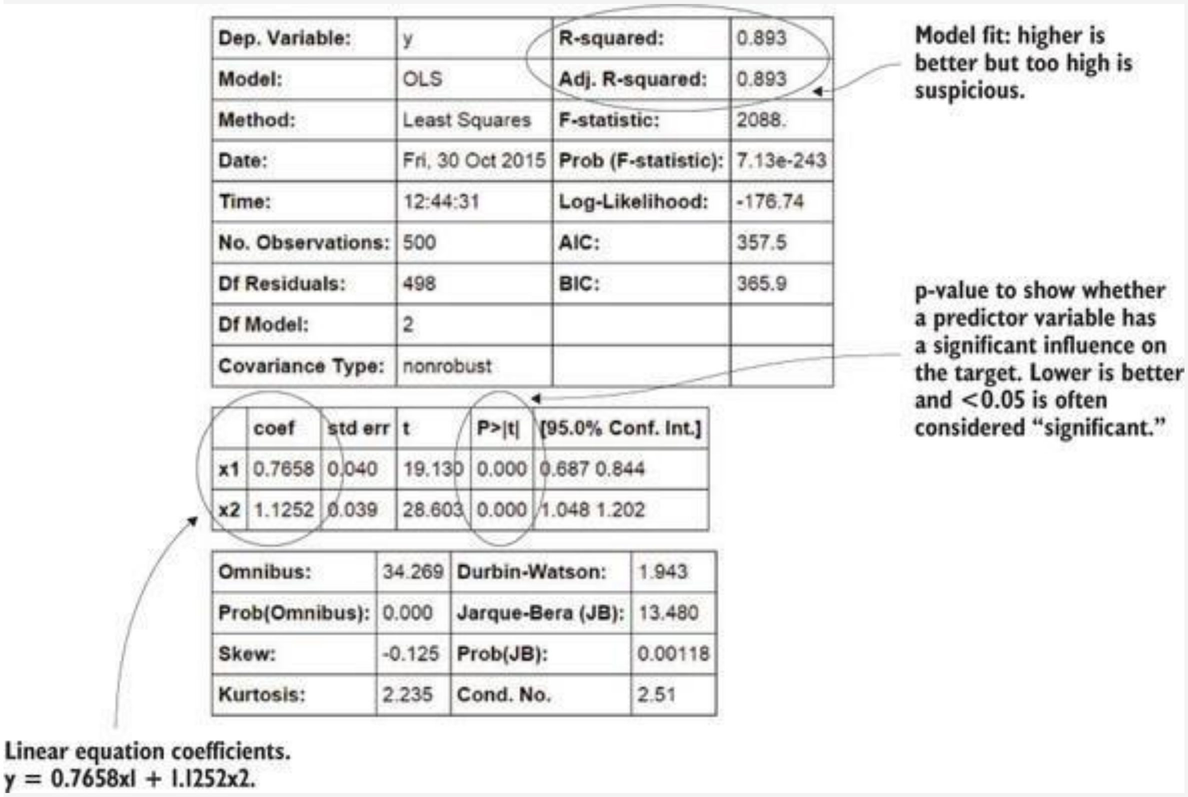
Shows model fit statistics.

Okay, we cheated here, quite heavily so. We created predictor values that are meant to predict how the target variables behave. For a linear regression, a “linear relation” between each x (predictor) and the y (target) variable is assumed, as shown in figure 22.

We, however, created the target variable, based on the predictor by adding a bit of randomness. It shouldn’t come as a surprise that this gives us a well-fitting model.

The `results.summary()` outputs the table in figure 23. Mind you, the exact outcome depends on the random variables you got.

Figure 23. Linear regression model information output



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.1252x2.

Let’s ignore most of the output we got here and focus on the most important parts:

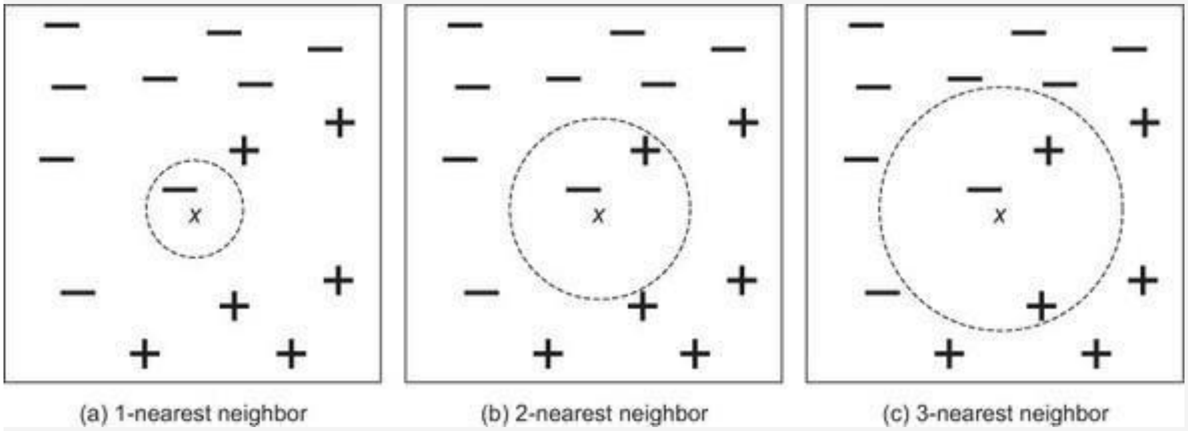
- ***Model fit*** —For this the R-squared or adjusted R-squared is used. This measure is an indication of the amount of variation in the data that gets captured by the model. The difference between the adjusted R-squared and the R-squared is minimal here because the adjusted one is the normal one + a penalty for model complexity. A model gets complex when many variables (or features) are introduced. You don’t need a complex model if a simple model is available, so the adjusted R-squared punishes you for overcomplicating. At any rate, 0.893 is high, and it should be because we cheated. Rules of thumb exist, but for models in businesses, models above 0.85 are often considered good. If you want to win a competition you need in the high 90s. For research however, often very low model fits (<0.2 even) are found. What's more important there is the influence of the introduced predictor variables.

- ***Predictor variables have a coefficient*** —For a linear model this is easy to interpret. In our example if you add "1" to x1, it will change y by "0.7658". It's easy to see how finding a good predictor can be your route to a Nobel Prize even though your model as a whole is rubbish. If, for instance, you determine that a certain gene is significant as a cause for cancer, this is important knowledge, even if that gene in itself doesn't determine whether a person will get cancer. The example here is classification, not regression, but the point remains the same: detecting influences is more important in scientific studies than perfectly fitting models (not to mention more realistic). But when do we know a gene has that impact? This is called significance.
- ***Predictor significance*** —Coefficients are great, but sometimes not enough evidence exists to show that the influence is there. This is what the p-value is about. A long explanation about type 1 and type 2 mistakes is possible here but the short explanations would be: if the p-value is lower than 0.05, the variable is considered significant for most people. In truth, this is an arbitrary number. It means there's a 5% chance the predictor doesn't have any influence. Do you accept this 5% chance to be wrong? That's up to you. Several people introduced the extremely significant (p<0.01) and marginally significant thresholds (p<0.1).

Linear regression works if you want to predict a value, but what if you want to classify something? Then you go to classification models, the best known among them being k-nearest neighbors.

As shown in figure 24, k-nearest neighbors looks at labeled points nearby an unlabeled point and, based on this, makes a prediction of what the label should be.

Figure 24. K-nearest neighbor techniques look at the k-nearest point to make a prediction.



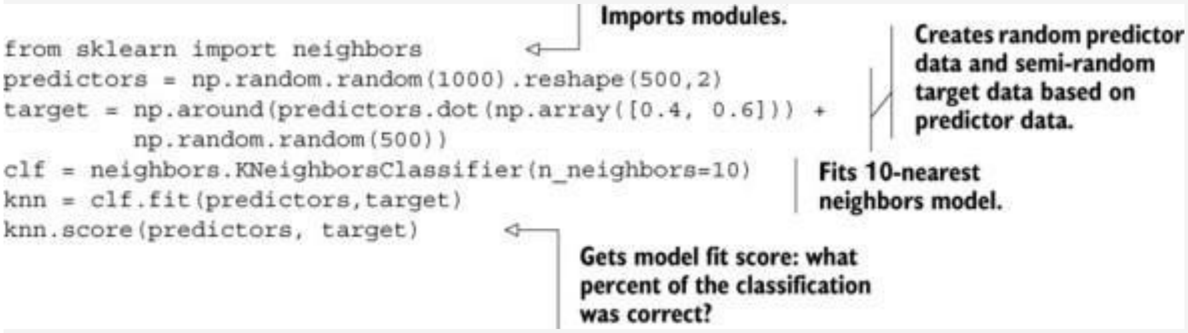
(a) 1-nearest neighbor

(b) 2-nearest neighbor

(c) 3-nearest neighbor

Let’s try it in Python code using the Scikit learn library, as in this next listing.

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?

As before, we construct random correlated data and surprise, surprise we get 85% of cases correctly classified. If we want to look in depth, we need to score the model. Don’t let `knn.score()` fool you; it returns the model accuracy, but by “scoring a model” we often mean applying it on data to make a prediction.

```
prediction = knn.predict(predictors)
```

copy

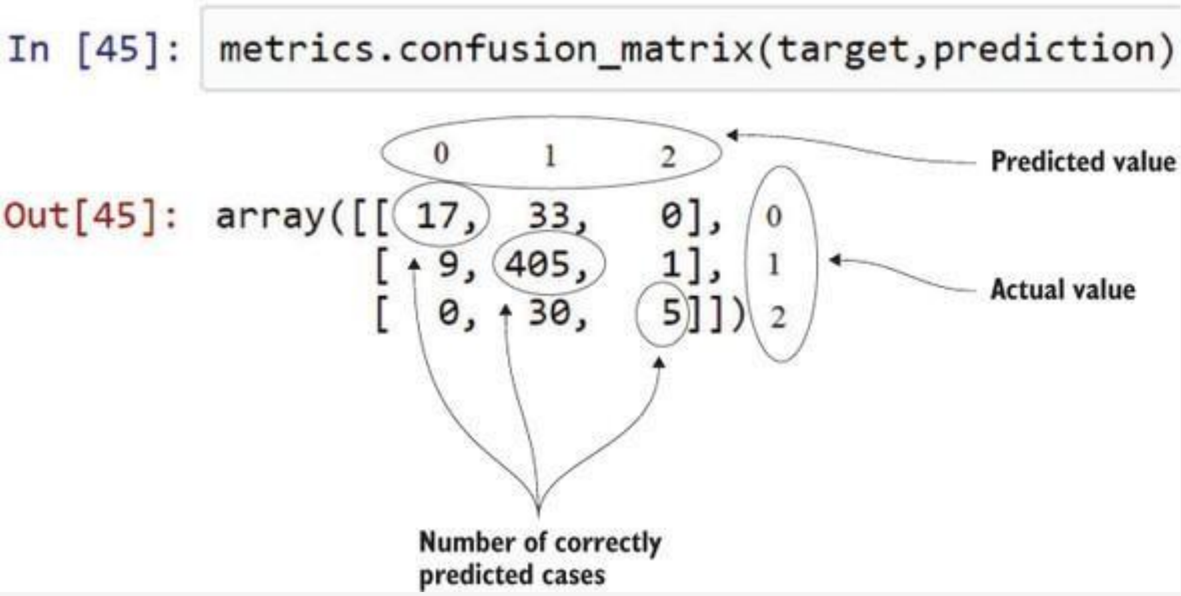
Now we can use the prediction and compare it to the real thing using a confusion matrix.

```
metrics.confusion_matrix(target,prediction)
```

copy

We get a 3-by-3 matrix as shown in figure 25.

Figure 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.



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

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

The confusion matrix shows we have correctly predicted 17+405+5 cases, so that’s good. But is it really a surprise? No, for the following reasons:

- For one, the classifier had but three options; marking the difference with last time `np.around()` will round the data to its nearest integer. In this case that’s either 0, 1, or 2. With only 3 options, you can’t do much worse than 33% correct on 500 guesses, even for a real random distribution like flipping a coin.
- Second, we cheated again, correlating the response variable with the predictors. Because of the way we did this, we get most observations being a “1”. By guessing “1” for every case we’d already have a similar result.
- We compared the prediction with the real values, true, but we never predicted based on fresh data. The prediction was done using the same data as the data used to build the

model. This is all fine and dandy to make yourself feel good, but it gives you no indication of whether your model will work when it encounters truly new data. For this we need a holdout sample, as will be discussed in the next section.

3. Model diagnostics and model comparison

You’ll be building multiple models from which you then choose the best one based on multiple criteria. Working with a holdout sample helps you pick the best-performing model. A holdout sample is a part of the data you leave out of the model building so it can be used to evaluate the model afterward. The principle here is simple: the model should work on unseen data. You use only a fraction of your data to estimate the model and the other part, the holdout sample, is kept out of the equation. The model is then unleashed on the unseen data and error measures are calculated to evaluate it. Multiple error measures are available, and in figure 2.26 we show the general idea on comparing models. The error measure used in the example is the mean square error.

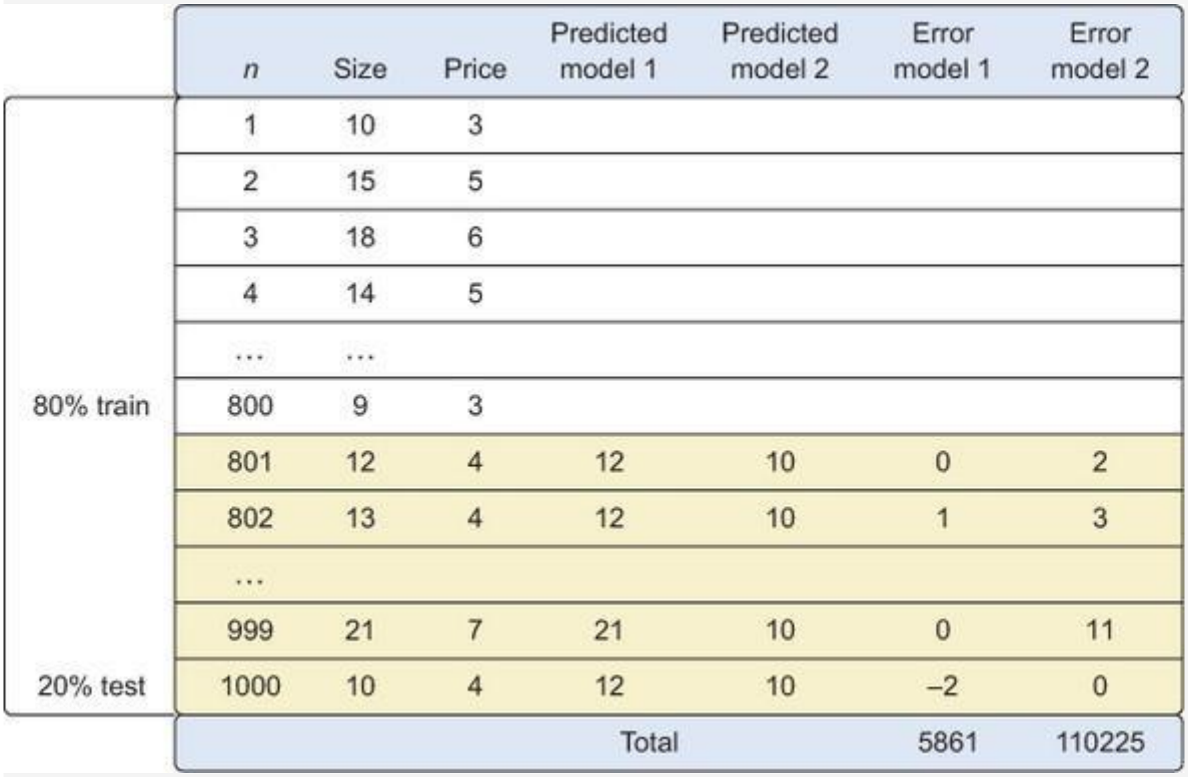
Figure 2.26. Formula for mean square error

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

Mean square error is a simple measure: check for every prediction how far it was from the truth, square this error, and add up the error of every prediction.

Figure 27 compares the performance of two models to predict the order size from the price. The first model is *size = 3 * price* and the second model is *size = 10*. To estimate the models, we use 800 randomly chosen observations out of 1,000 (or 80%), without showing the other 20% of data to the model. Once the model is trained, we predict the values for the other 20% of the variables based on those for which we already know the true value, and calculate the model error with an error measure. Then we choose the model with the lowest error. In this example we chose model 1 because it has the lowest total error.

Figure 27. A holdout sample helps you compare models and ensures that you can generalize results to data that the model has not yet seen.



	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

Many models make strong assumptions, such as independence of the inputs, and you have to verify that these assumptions are indeed met. This is called *model diagnostics*.

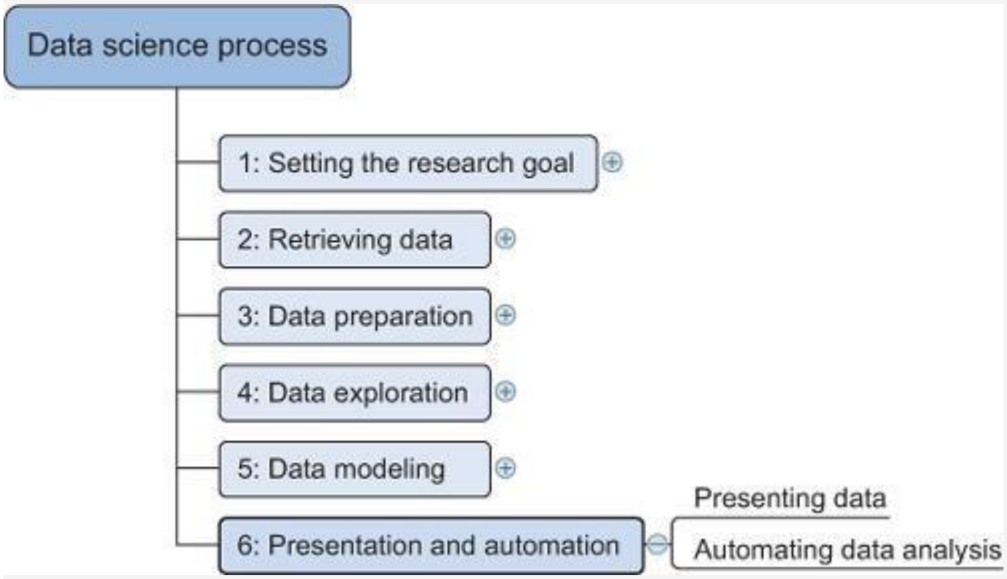
This section gave a short introduction to the steps required to build a valid model. Once you have a working model you're ready to go to the last step.

10) Presentation and automation

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

After you've successfully analyzed the data and built a well-performing model, you're ready to present your findings to the world (figure 28). This is an exciting part; all your hours of hard work have paid off and you can explain what you found to the stakeholders.

Figure 28. Step 6: Presentation and automation



Sometimes people get so excited about your work that you'll need to repeat it over and over again because they value the predictions of your models or the insights that you produced. For this reason, you need to automate your models. This doesn't always mean that you have to redo all of your analysis all the time. Sometimes it's sufficient that you implement only the model scoring; other times you might build an application that automatically updates reports, Excel spreadsheets, or PowerPoint presentations. The last stage of the data science process is where your *soft skills* will be most useful, and yes, they're extremely important. In fact, we recommendyou find dedicated books and other information on the subject and work through them, because why bother doing all this tough work if nobody listens to what you have to say?

If you've done this right, you now have a working model and satisfied stakeholders, so we canconclude this chapter here.

Summary

the data science process consists of six steps:

- ***Setting the research goal*** —Defining the what, the why, and the how of your project in a project charter.
- ***Retrieving data*** —Finding and getting access to data needed in your project. This data is either found within the company or retrieved from a third party.
- ***Data preparation*** —Checking and remediating data errors, enriching the data with data from other data sources, and transforming it into a suitable format for your models.
- ***Data exploration*** —Diving deeper into your data using descriptive statistics and visual techniques.
- ***Data modeling*** —Using machine learning and statistical techniques to achieve your project goal.
- ***Presentation and automation*** —Presenting your results to the stakeholders and industrializing your analysis process for repetitive reuse and integration with other tools.

D3491 FUNDAMENTALS OF DATA SCIENCE

UNIT 2

FrequencyDistributionandData:Types,Tables,andGraphs

Frequency distribution in statistics provides the information of the number of occurrences (frequency) of distinct values distributed within a given period of time or interval, in a list, table,or graphical representation.

Types of Frequency Distribution:

There are two types of Frequency Distribution.

- Grouped
- Ungrouped

There are two types Data is a **collection of numbers** or **values**

Data: Any bit of information that is expressed in a **value or numerical number** is data. Data isbasically a collection of information, measurements or observations.

For example

- The marks you scored in your Math exam is data
- The number of cars that pass through a bridge in a day.

Raw data :

Raw data is an initial collection of information. This information has not yet been organized. Afterthe very first step of data collection, you will get raw data. For example,

A group of five friends their favourite colour. The answers are Blue, Green, Blue, Red, and Red. Thiscollection of information is the raw data.

Discrete data :***Discrete data*** is that which is recorded in whole numbers, like the number ofchildren in a school or number of tigers in a zoo. It cannot be in decimals or fractions.

Continuous data :***Continuous data*** need not be in whole numbers, it can be in decimals. Examplesare the temperature in a city for a week, your percentage of marks for the last exam etc.

Example of Data Handling:

- Pictographs

- Bar Graphs
- Histogram and Pie-Charts
- Chance and Probability
- Arithmetic Mean and Median and Mode

Frequency

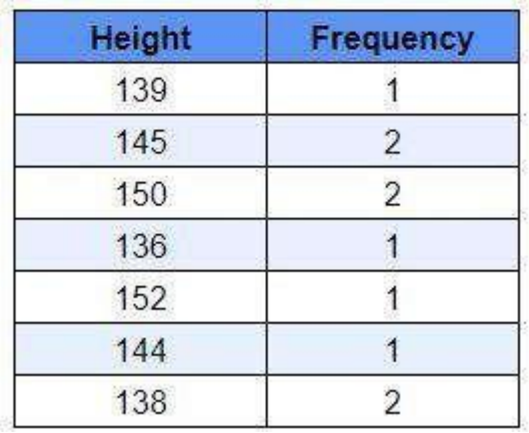
The frequency of any value is the number of times that value appears in a data set. So from the above examples of colours, we can say two children like the colour blue, so its frequency is two. So to make meaning of the raw data, we must organize. And finding out the frequency of the data values is how this organisation is done.

Frequency Distribution

Many times it is not easy or feasible to find the frequency of data from a very large dataset. So to make sense of the data we make a frequency table and graphs. Let us take the example of the heights of ten students in cms.

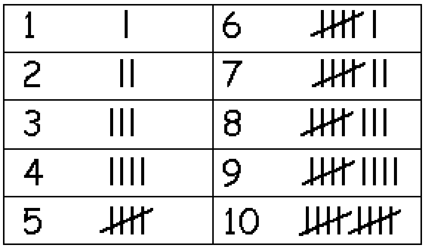
Frequency Distribution Table

139, 145, 150, 145, 136, 150, 152, 144, 138, 138



Height	Frequency
139	1
145	2
150	2
136	1
152	1
144	1
138	2

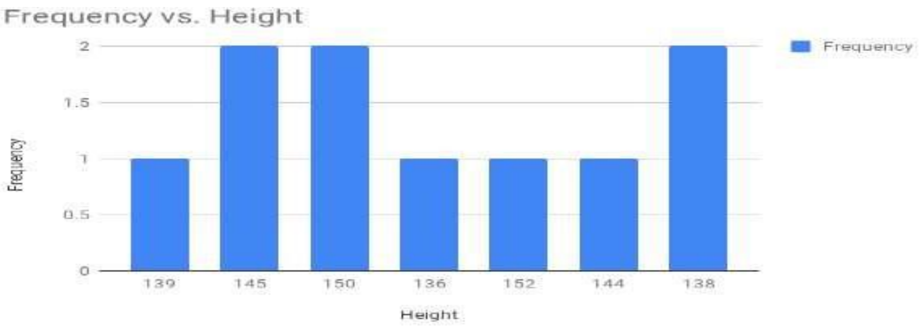
This frequency table will help us make better sense of the data given. Also when the data set is too big (say if we were dealing with 100 students) we use tally marks for counting. It makes the task more organized and easy. Below is an example of how we use tally marks.



1	𝍷	6	𝍸𝍷
2	𝍷𝍷	7	𝍸𝍷𝍷
3	𝍷𝍷𝍷	8	𝍸𝍷𝍷𝍷
4	𝍷𝍷𝍷𝍷	9	𝍸𝍷𝍷𝍷𝍷
5	𝍸	10	𝍸𝍸

Frequency Distribution Graph

Using the same above example we can make the following graph:



Learn more about Bar Graphs and Histogram here.

Types of Frequency Distribution

- Grouped frequency distribution.
- Ungrouped frequency distribution.
- Cumulative frequency distribution.
- Relative frequency distribution.
- Relative cumulative frequency distribution.

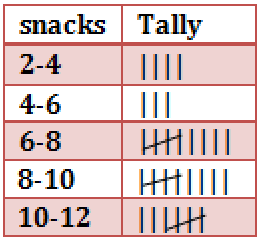
Grouped Data

At certain times to ensure that we are making correct and relevant observations from the data set, we may need to group the data into class intervals. This ensures that the frequency distribution best represents the data. example :the height of students.

Class Interval	Frequency
130-140	4
140-150	3
150-160	3

From the above table, you can see that the value of 150 is put in the class interval of 150-160 and not 140-150. This is the convention we must follow.

- **The table gives the number of snacks ordered and the number of days as a tally. Find the frequency of snacks ordered.** **2**



snacks	Tally
2-4	||||
4-6	|||
6-8	卌||||
8-10	卌||||
10-12	||卌

Answer: From the frequency table the number of snacks ordered ranging between

- 2-4 is 4 days
- 4 to 6 is 3 days
- 6 to 8 is 9 days
- 8 to 10 is 9 days
- 10 to 12 is 7 days.

So the frequencies for all snacks ordered are 4, 3, 9, 9, 7

- **How to find frequency distribution?** **2**

Answer: We can find frequency distribution by the following steps:

- First of all, calculate the range of the data set.
- Next, divide the range by the number of the group you want your data in and then round up.
- After that, use class width to create groups
- Finally, find the frequency for each group.

- **Define frequency distribution in statistics?** **2**

Answer: In an overview, the frequency distribution of all distinct values in some variables and the number of times they occur. Meaning that it tells how frequencies are distributed overvalues in a frequency distribution. However, mostly we use frequency distributions to summarize categorical variables.

- **Why are frequency distributions important? 2**

Answer: It has great importance in statistics. Also, a well-structured frequency distribution makes possible a detailed analysis of the structure of the population with respect to given characteristics. Therefore, the groups into which the population break down can be determined.

- **State the components of frequency distribution? 2**

Answer: The various components of the frequency distribution are: Class interval, types of class interval, class boundaries, midpoint or class mark, width or size o class interval, class frequency,

frequency density = class frequency/ class width,

relative frequency = class frequency/ total frequency, etc.

Descriptive Statistics

A population is the group to be studied, and population data is a collection of all elements in the population. For example:

- All the fish in Long Lake.
- All the lakes in the Adirondack Park.
- All the grizzly bears in Yellowstone National Park.

A sample is a subset of data drawn from the population of interest. For example:

- 100 fish randomly sampled from Long Lake.
- 25 lakes randomly selected from the Adirondack Park.
- 60 grizzly bears with a home range in Yellowstone National Park.

Populations are characterized by descriptive measures called parameters. Inferences about parameters are based on sample statistics.

For example,

The population mean (μ) is estimated by the sample mean ($\bar{x}$). The population variance (σ2) is estimated by the sample variance (s2).

Variables are the characteristics we are interested in.

For example:

- The length of fish in Long Lake.
- The pH of lakes in the Adirondack Park.

- The weight of grizzly bears in Yellowstone National Park.

Variables are divided into two major groups: **Qualitative And Quantitative**.

1. **Qualitative variables**

- Qualitative variables have values that are attributes or categories.
- Mathematical operations cannot be applied to qualitative variables.
- Examples of qualitative variables are gender, race, and petal color.
- Quantitative variables have values that are typically numeric, such as measurements.
- Mathematical operations can be applied to these data. Examples of quantitative variables are age, height, and length.

2. **Quantitative variables**

    - Quantitative variables can be broken down further into two more categories: discrete and continuous variables.
    - **Discrete variables** have a finite or countable number of possible values. Think of discrete variables as “hens.” Hens can lay 1 egg, or 2 eggs, or 13 eggs… There are a limited, definable number of values that the variable could take on.
    - **Continuous variables** have an infinite number of possible values. Think of continuous variables as “cows.” Cows can give 4.6713245 gallons of milk, or 7.0918754 gallons of milk, or 13.272698 gallons of milk … There are an almost infinite number of values that a continuous variable could take on.

Examples

Is the variable qualitative or quantitative?

Species	Weight	Diameter	Zip Code
(qualitative	quantitative,	quantitative,	qualitative)

Graphs

Data can be described clearly and concisely with the aid of a well-constructed frequency distribution.

GRAPHS FOR QUANTITATIVE DATA

Histograms

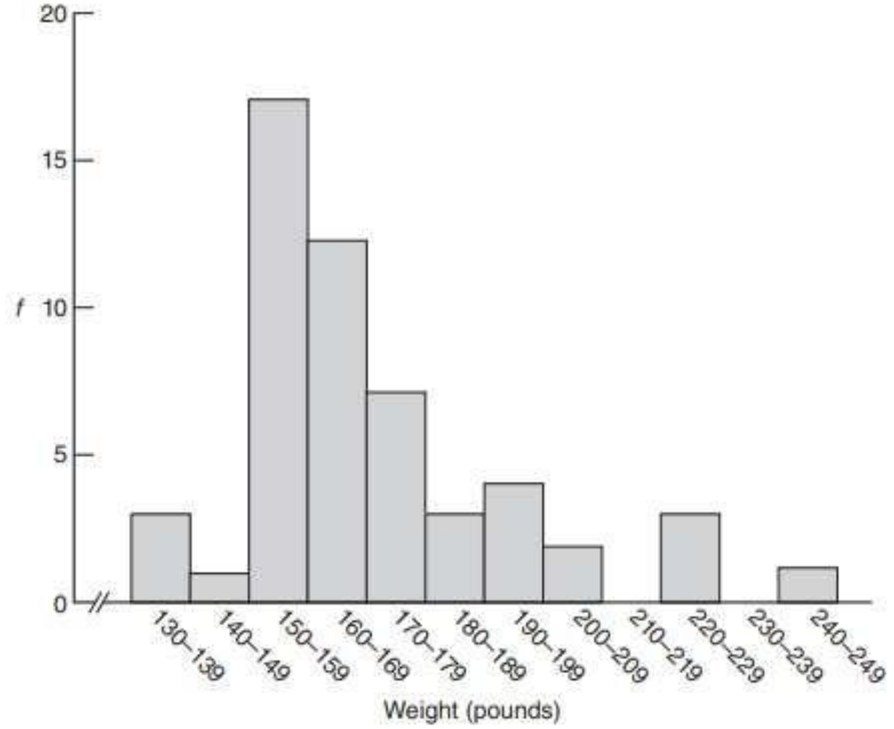
A bar-type graph for quantitative data. The common boundaries between adjacent bars emphasize the continuity of the data, as with continuous variables.

A histogram in Figure shows a casual glance at this histogram confirms previous conclusions: a dense concentration of weights among the 150s, 160s, and 170s, with a spread in the direction of the heavier weights. Let's pinpoint some of the more important features of histograms.

■ Equal units along the horizontal axis (the X axis, or abscissa) reflect the various class intervals of the frequency distribution.

■ Equal units along the vertical axis (the Y axis, or ordinate) reflect increases in frequency. (The units along the vertical axis do not have to be the same width as those along the horizontal axis.)

■ The intersection of the two axes defines the origin at which both numerical scales equal 0



Frequency Polygon

A line graph for quantitative data that also emphasizes the continuity of continuous variables

An important variation on a histogram is the frequency polygon, or line graph. Frequency polygons may be constructed directly from frequency distributions. However, we will follow the step-by-step transformation of a histogram into a frequency polygon, as described in panels A, B, C, and D of Figure 2.2. A. This panel shows the histogram for the weight distribution. B. Place dots at the midpoints of each bar top or, in the absence of bar tops, at midpoints for classes on the horizontal axis, and connect them with straight lines. [To find the midpoint of any class, such as 160–169, simply add the two tabled boundaries (160 + 169 = 329) and divide this sum by 2 (329/2 = 164.5).] C. Anchor the frequency polygon to the horizontal axis. First, extend the upper tail to the midpoint of the first unoccupied class (250–259) on the upper flank of the histogram. Then extend the lower tail to the midpoint of the first unoccupied class (120–129) on the lower flank of the histogram. Now all of the area under the frequency polygon is enclosed completely. D. Finally, erase all of the histogram bars, leaving only the frequency polygon. Frequency polygons are particularly useful when two or more frequency distributions or relative frequency distributions are to be included in the same graph.

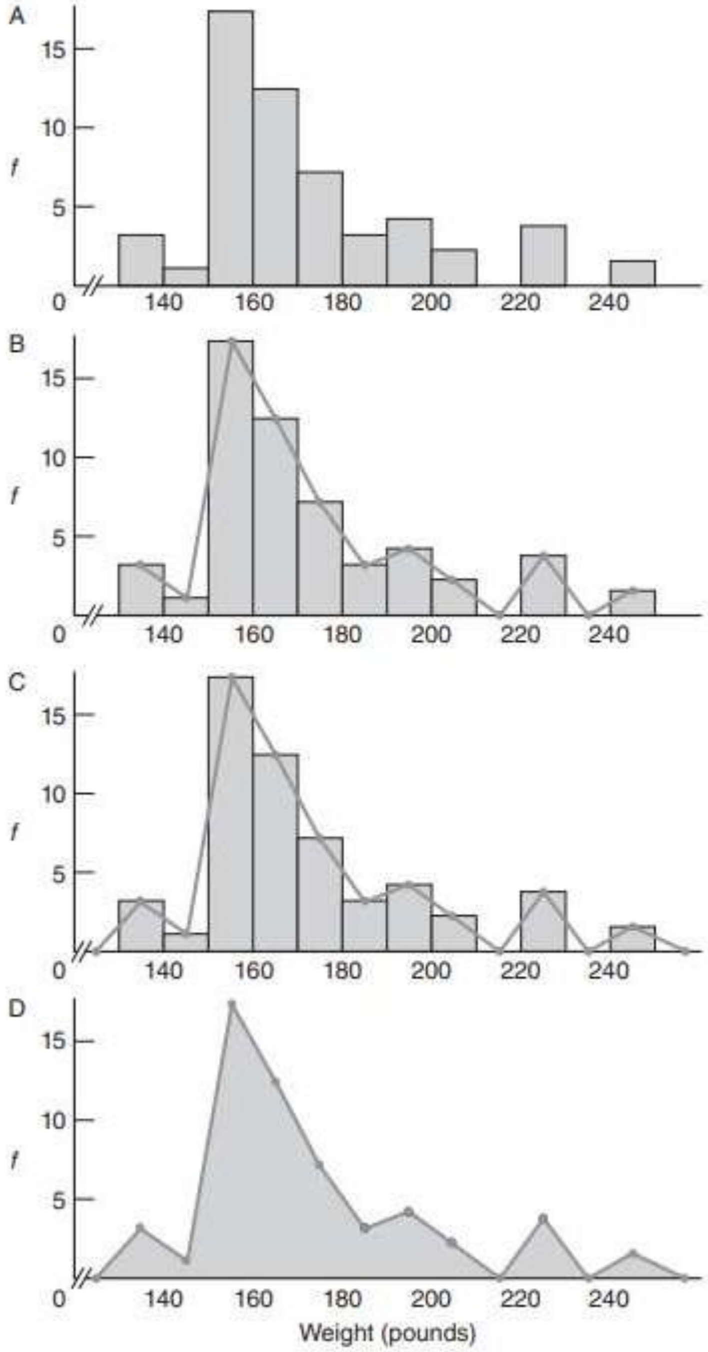


FIGURE 2.2

Transition from histogram to frequency polygon.

Stem and Leaf Displays:

A device for sorting quantitative data on the basis of leading and trailing digits.

Still another technique for summarizing quantitative data is a stem and leaf display. Stem and leaf displays are ideal for summarizing distributions, such as that for weight data, without destroying the identities of individual observations.

Constructing a Display

The stemplot (also called stem and leaf plot) is another graphical display ofthe distribution of quantitative variable.

To create a stemplot, the idea is to separate each data point into a stemand leaf, asfollows:

- The leaf is the right-most digit.
- The stem is everything except the right-most digit.
- So, if the data point is 34, then 3 is the stem and 4 is the leaf.
- If the data point is 3.41, then 3.4 is the stem and 1 is the leaf.
- Note: For this to work, ALL data points should be rounded to the same number of decimal places.

EXAMPLE: Best Actress Oscar Winners

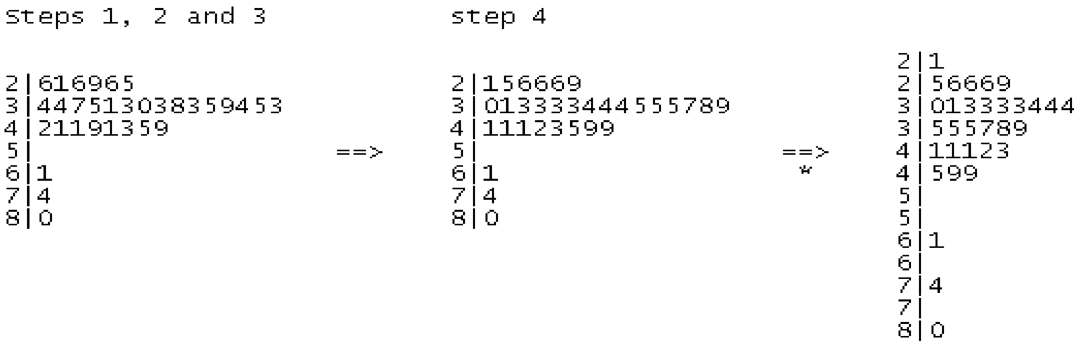
We will continue with the Best Actress Oscar winners example

34 34 26 37 42 41 35 31 41 33 30 74 33 49 38 61 21 41 26 80 43 29 33 35 45
49 39 34 26 25 35 33

To make a stemplot:

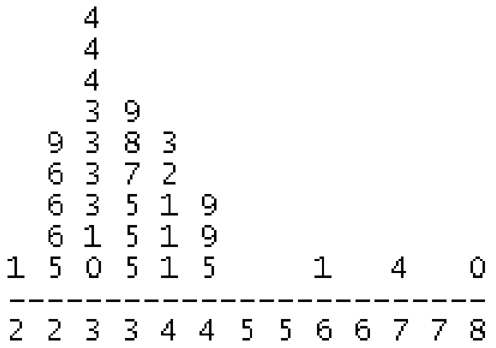
- Separate each observation into a stem and a leaf.
- Write the stems in a vertical column with the smallest at the top, and draw avertical line at the right of this column.
- Go through the data points, and write each leaf in the row to the right of itsstem.
- Rearrange the leaves in an increasing order.

When some of the stems hold a large number of leaves, we can split each stem into two: one holding the leaves 0-4, and the other holding the leaves 5-9. A



```
Steps 1, 2 and 3              step 4
                                                                   2|1
2|616965                      2|156669                             2|56669
3|447513038359453             3|013333444555789                    3|013333444
4|21191359                    4|11123599                           3|555789
5|                    ==>     5|                           ==>     4|11123
6|1                           6|1                           *      4|599
7|4                           7|4                                  5|
8|0                           8|0                                  5|
                                                                   6|1
                                                                   6|
                                                                   7|4
                                                                   7|
                                                                   8|0
```

statistical software package will often do the splitting for you, when appropriate.Note that when rotated 90 degrees counter-clockwise, the stemplot visuallyresembles a histogram:



```
    4
    4
    4
    3 9
  9 3 8 3
  6 3 7 2
  6 3 5 1 9
  6 1 5 1 9
1 5 0 5 1 5     1   4   0
-------------------------
2 2 3 3 4 4 5 5 6 6 7 7 8
```

The stemplot has additional unique features:

- preserves the original data.
- It sorts the data (which will become very useful in the next section).

Typical Shapes

Whether expressed as a histogram, a frequency polygon, or a stem and leaf display, an important characteristic of a frequency distribution is its shape. Figure 2.3 shows some of the more typical shapes for smoothed frequency polygons (which ignore the inevitable irregularities of real data).

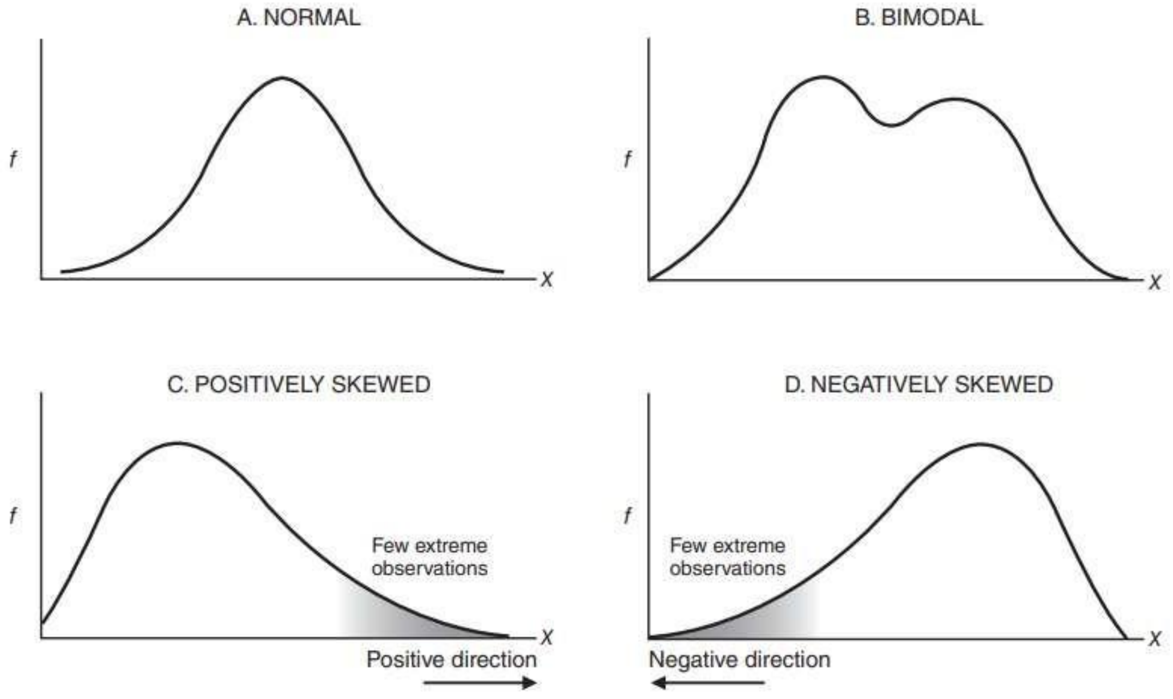


FIGURE 2.3
Typical shapes.

Normal

Any distribution that approximates the normal shape in panel A of Figure 2.3 can be analyzed, as we will see in Chapter 5, with the aid of the well-documented normal curve. The familiar bell-shaped silhouette of the normal curve can be superimposed on many frequency distributions, including those for uninterrupted gestation periods of human fetuses, scores on standardized tests, and even the popping times of individual kernels in a batch of popcorn.

Bimodal

Any distribution that approximates the bimodal shape in panel B of Figure 2.3 might, as suggested previously, reflect the coexistence of two different types of observations in the same distribution. For instance, the distribution of the ages of residents in a neighborhood consisting largely of either new parents or their infants has a bimodal shape.

Positively Skewed The two remaining shapes in Figure 2.3 are lopsided. A lopsided distribution caused by a few extreme observations in the positive direction (to the right of the majority of observations), as in panel C of Figure 2.3, is a positively skewed distribution.

The distribution of incomes among U.S. families has a pronounced positive skew, with most family incomes under $200,000 and relatively few family incomes spanning a wide range of values above $200,000. The distribution of weights in Figure 2.1 also is positively skewed.

Negatively Skewed A lopsided distribution caused by a few extreme observations in the negative direction (to the left of the majority of observations), as in panel D of Figure 2.3, is a negatively skewed distribution. The distribution of ages at retirement among U.S. job holders has a pronounced negative skew, with most retirement ages at 60 years or older and relatively few retirement ages spanning the wide range of ages younger than 60.

A GRAPH FOR QUALITATIVE (NOMINAL) DATA:

The distribution in Table 2.7, based on replies to the question "Do you have a Facebook profile?" appears as a bar graph in Figure 2.4. A glance at this graph confirms that Yes replies occur approximately twice as often as No replies. As with histograms, equal segments along the horizontal axis are allocated to the different words or classes that appear in the frequency distribution for qualitative data. Likewise, equal segments along the vertical axis reflect increases in frequency. The body of the bar graph consists of a series of bars whose heights reflect the frequencies for the various words or classes. A person's answer to the question "Do you have a Facebook profile?" is either Yes or No, not some impossible intermediate value, such as 40 percent Yes and 60 percent No. Gaps are placed between adjacent bars of bar graphs to emphasize the discontinuous nature of qualitative data. A bar graph also can be used with quantitative data to emphasize the discontinuous nature of a discrete variable, such as the number of children in a family.

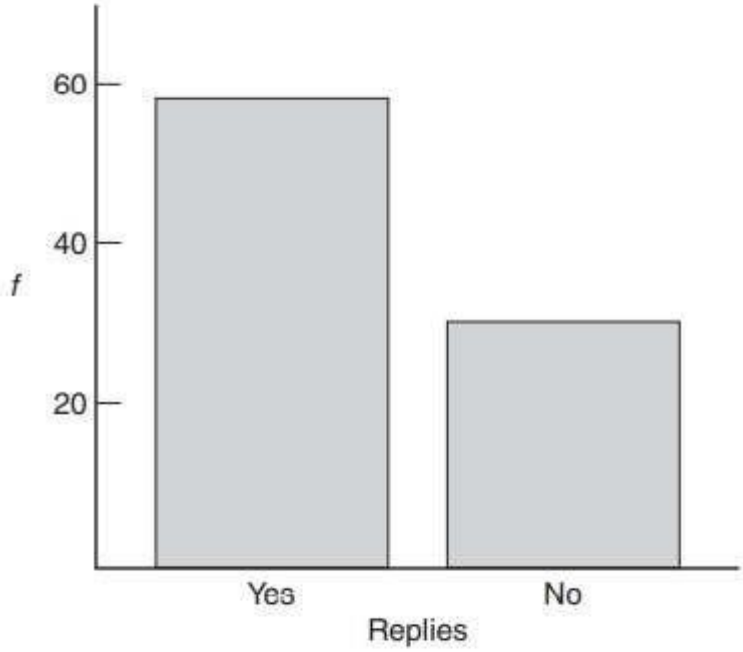


FIGURE 2.4
Bar graph.

Misleading Graphs:

Graphs can be constructed in an unscrupulous manner to support a particular point of view. Indeed, this type of statistical fraud gives credibility to popular sayings, including "Numbers don't lie, but statisticians do" and "There are three kinds of lies—lies, damned lies, and statistics." For example, to imply that comparatively many students responded Yes to the Facebook profile question, an unscrupulous person might resort to the various tricks shown in Figure 2.5:

■ The width of the Yes bar is more than three times that of the No bar, thus violating the custom that bars be equal in width.
■ The lower end of the frequency scale is omitted, thus violating the custom that the entire scale be reproduced, beginning with zero. (Otherwise, a broken scale should be highlighted by crossover lines, as in Figures 2.1 and 2.2.)
■ The height of the vertical axis is several times the width of the horizontal axis, thus violating the custom, heretofore unmentioned, that the vertical axis be approximately as tall as the horizontal axis is wide. Beware of graphs in which, because the vertical axis is many times larger than the horizontal axis (as in Figure 2.5), frequency differences are exaggerated, or in which, because the vertical axis is many times smaller than the horizontal axis, frequency differences are suppressed.

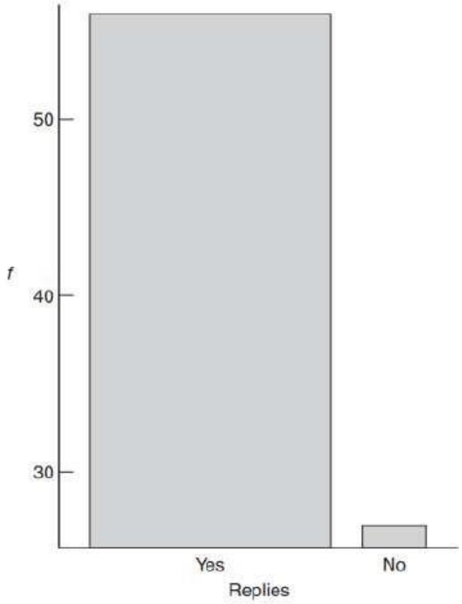


FIGURE 2.5
Distorted bar graph.

AVERAGES

A center of a data set is a way of describing a location. We can measure a center of a data in 3 different ways: the mean (average), the median and the mode.

The two main numerical measures for the center of a distribution are the mean and the median. Each one of these measures is based on a completely different idea of describing the center of a distribution. Let us first present each one of the measures, and then compare their properties.

MEAN

The mean is the average of a set of observations (i.e., the sum of the observations divided by the number of observations).

The mean is the average of a set of observations. If the n observations are written as their mean can be written mathematically as: their mean is:

We read the symbol as "x-bar." The bar notation is commonly used to represent the samplemean, i.e. the mean of the sample.

EXAMPLE: Best Actress Oscar Winners

We will continue with the Best Actress Oscar winners example .

34 34 26 37 42 41 35 31 41 33 30 74 33 49 38 61 21 41 26 80 43 29 33 35 45

49 39 34 26 25 35 33

The mean age of the 32 actresses is:

We add all of the ages to get 1233 and divide by the number of ages which was 32 to get 38.5. We denote this result as x-bar and called the sample mean.

EXAMPLE: World Cup Soccer

Often we have large sets of data and use a frequency table to display the data more efficiently. Data were collected from the last three World Cup soccer tournaments. A total of 192 games were played. The table below lists the number of goals scored per game (not including any goals scored in shootouts).

Total # Goals/Game	**Frequency**
0	17
1	45
2	51

3	37
4	25
5	11
6	3
7	2
8	1

To find the mean number of goals scored per game, we would need to find the sum of all 192 numbers, and then divide that sum by 192.

Rather than add 192 numbers, we use the fact that the same numbers appear many times. For example, the number 0 appears 17 times, the number 1 appears 45 times, the number2 appears 51 times, etc.

If we add up 17 zeros, we get 0. If we add up 45 ones, we get 45. If we add up 51 twos, we get 102. Repeated addition is multiplication.

Thus, the sum of the 192 numbers

= 0(17) + 1(45) + 2(51) + 3(37) + 4(25) + 5(11) + 6(3) + 7(2) + 8(1) = 453.

The sample mean is then 453 / 192 = 2.359.

Note that, in this example, the values of 1, 2, and 3 are the most common andour averagefalls in this range representing the bulk of the data.

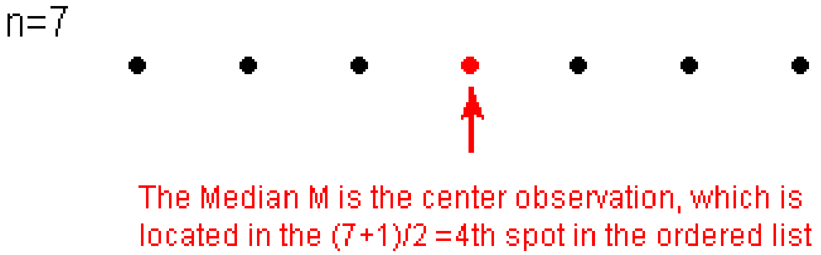
MEDIAN

Define and calculate the sample median of a quantitative variable.
The median M is the midpoint of the distribution. It is the number suchthat half of the observations fall above, and half fall below.
To find the median:
Order the data from smallest to largest.
Consider whether n, the number of observations, is even or odd.
If n is odd, the median M is the center observation in the ordered list. Thisobservation is the one “sitting” in the (n + 1) / 2 spot in the ordered list.

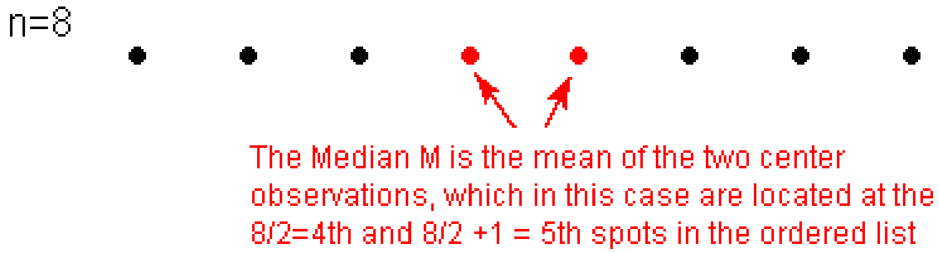
If n is even, the median M is the mean of the two center observations in the ordered list. These two observations are the ones “sitting” in the (n / 2) and(n / 2) + 1 spots in the ordered list.

EXAMPLE: Median (1)

For a simple visualization of the location of the median, consider the followingtwo simple cases of n = 7 and n = 8 ordered observations, with each observation represented by asolid circle:



The Median M is the center observation, which is located in the (7+1)/2 =4th spot in the ordered list



The Median M is the mean of the two center observations, which in this case are located at the 8/2=4th and 8/2 +1 = 5th spots in the ordered list

Comments:

In the images above, the dots are equally spaced, this need not indicate the data values are actually equally spaced as we are only interested in listing them in order. In fact, in the above pictures, two subsequent dots could have exactly the same value. It is clear that the value of the median will be in the same position regardless of the distance between data values.

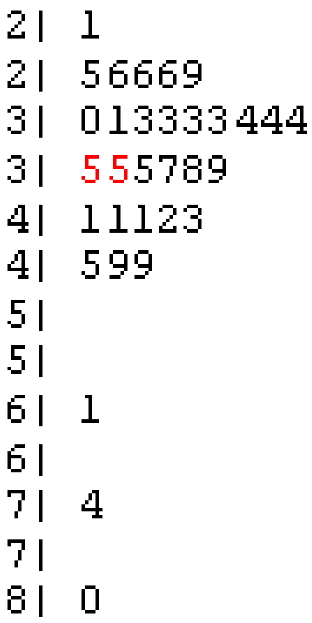
EXAMPLE: Median (2)

To find the median age of the Best Actress Oscar winners, we first need to order the data. It would be useful, then, to use the stemplot, a diagram in which the dataare already ordered.

Here n = 32 (an even number), so the median M, will be the mean of thetwo center observations.

These are located at the (n / 2) = 32 / 2 = 16th and(n / 2) +
1 = (32 / 2) + 1 = 17th

Counting from the top, we find that: the 16th ranked observation is 35the 17thranked observation also happens to be 35. Therefore, the median M = (35 + 35) / 2 = 35



```
2| 1
2| 56669
3| 013333444
3| 555789
4| 11123
4| 599
5|
5|
6| 1
6|
7| 4
7|
8| 0
```

Comparing the Mean and the Median

The mean and the median, the most common measures of center, each describe the centerof a distribution of values in a different way.

The mean describes the center as an average value, in which the actual values of the data points play an important role.

The median, on the other hand, locates the middle value as the center, and theorder of the data is the key.

To get a deeper understanding of the differences between these twomeasures of center,consider the following example. Here are two datasets:

Data set A → 64 65 66 68 70 71 73

Data set B → 64 65 66 68 70 71 730

For dataset A, the mean is 68.1, and the median is 68.

Looking at dataset B, notice that all of the observations except the last one are close together. The observation 730 is very large, and is certainly an outlier. In this case, the median is still 68, but the mean will be influenced by the high outlier, and shifted up to 162.

The message that we should take from this example is:

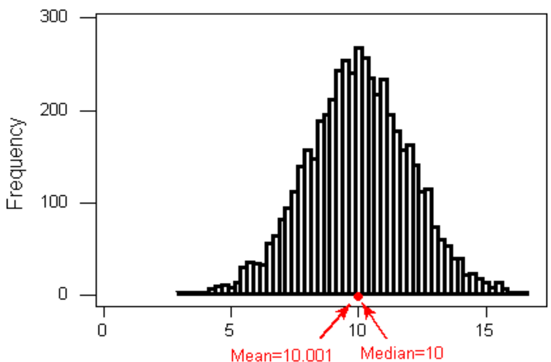
The mean is very sensitive to outliers (because it factors in their magnitude), while the median is resistant (or robust) to outliers.

MODE: 3rd Measure

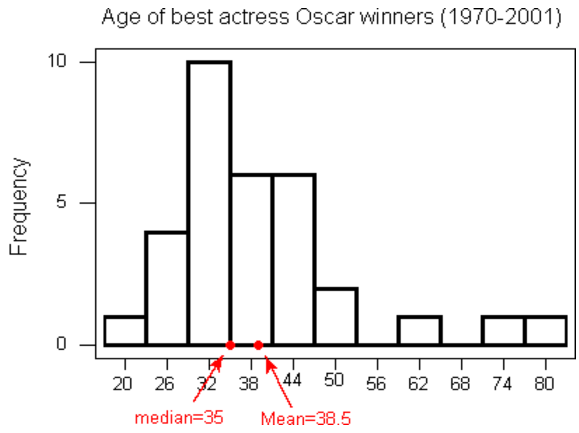
The mode of a data set is the number that occurs most frequently in the set.

- If no value appears more than once in the data set, the data set has no mode.
- If a there are two values that appear in the data set an equal number of times, theyboth will be modes etc.

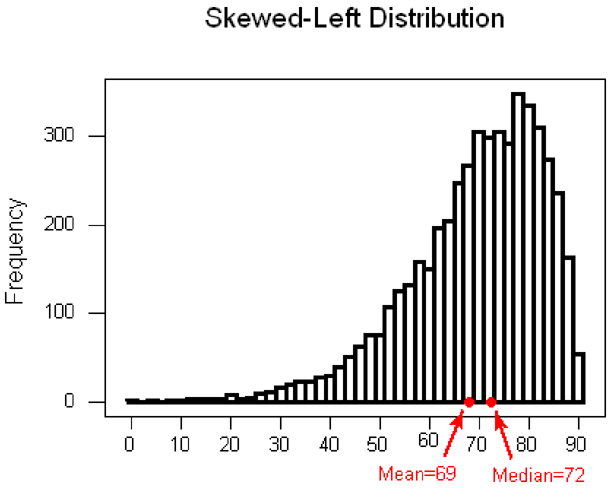
For symmetric distributions with no outliers: the mean is approximately equaltothe median.



For skewed right distributions and/or datasets with high outliers: the mean is greater than the median.



For skewed left distributions and/or datasets with low outliers: the mean is less than the median.



When to use which measures?

- Use the sample mean as a measure of center for symmetric distributions with no outliers. Otherwise, the median will be a more appropriate measure of the center of our data.

Let's Summarize

- The two main numerical measures for the center of a distribution are the mean and the median. The mean is the average value, while the median is the middle value.

- The mean is very sensitive to outliers (as it factors in their magnitude), while the median is resistant to outliers.
- The mean is an appropriate measure of center for symmetric distributions with no outliers. In all other cases, the median is often a better measure of the center of the distribution.

Describing Variability

Intuitive Approach

- In Figure 4.1, each of the three frequency distributions consists of seven scores with the same mean (10) but with different variabilities. (Ignore the numbers in boxes; their significance will be explained later.) Before reading on, rank the three distributions from least to most variable. Your intuition was correct if you concluded that distribution A has the least variability, distribution B has intermediate variability, and distribution C has the most variability. If this conclusion is not obvious, look at each of the three distributions, one at a time, and note any differences among the values of individual scores. For distribution A with the least (zero) variability, all seven scores have the same value (10). For distribution B with intermediate variability, the values of scores vary slightly (one 9 and one 11), and for distribution C with most variability, they vary even more (one 7, two 9s, two 11s, and one 13). Importance of Variability Variability assumes a key role in an analysis of research results. For example, a researcher might ask: Does fitness training improve, on average, the scores of depressed patients on a mental-wellness test? To answer this question, depressed patients are randomly assigned to two groups, fitness training is given to one group, and wellness scores are obtained for both groups. Let's assume that the mean wellness score is larger for the group with fitness training. Is the observed mean difference between the two groups real or merely transitory? This decision depends not only on the size of the mean difference between the two groups but also on the inevitable variabilities of individual scores within each group. To illustrate the importance of variability, Figure 4.2 shows the outcomes for two fictitious experiments, each with the same mean difference of 2, but with the two groups in experiment B having less variability than the two groups in experiment C. Notice that groups B and C in Figure 4.2 are the same as their counterparts in Figure 4.1. Although the new group B* retains exactly the same (intermediate) variability as group B, each of its seven scores and its mean have been shifted 2 units to the right. Likewise, although the new group C* retains exactly the same (most) variability as group C, each of its seven scores and its mean have been shifted 2 units to the right. Consequently, the crucial mean difference of 2 (from 12 − 10 = 2) is the same for both experiments. Before reading on, decide which mean difference of 2 in Figure 4.2 is more apparent. The mean difference for experiment B should seem more apparent because of the smaller variabilities within both groups B and B*. Just as it's easier to hear a phone message when static is reduced, it's easier to see a difference between group means when variabilities within groups are reduced.

DESCRIBING VARIABILITY

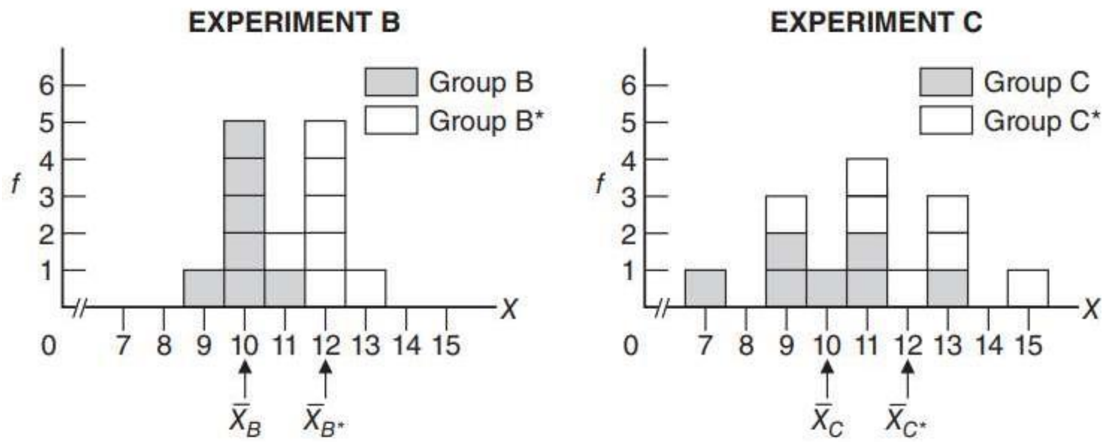


FIGURE 4.2

Two experiments with the same mean difference but dissimilar variabilities.

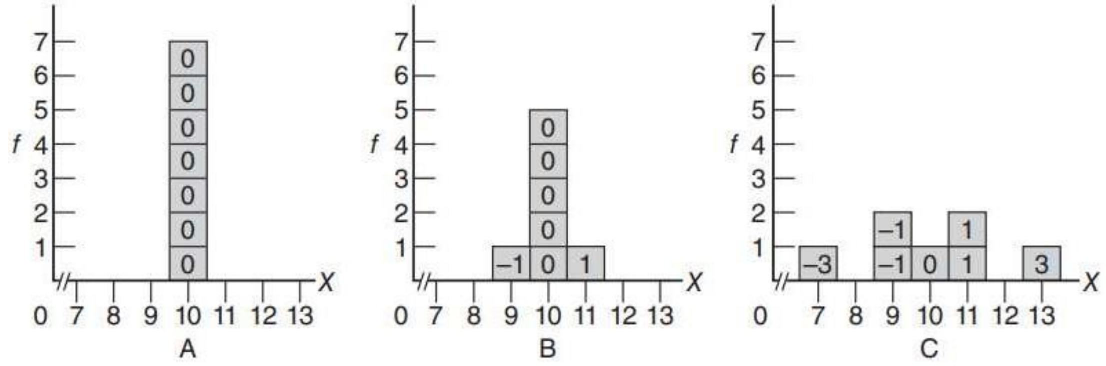


FIGURE 4.1

Three distributions with the same mean (10) but different amounts of variability. Numbers in the boxes indicate distances from the mean.

Range

A range measures the spread of a data inside the limits of a data set, it is calculated as a difference between the highest and lowest values in the data set. The larger the range, the greater the spread of the data.The range covered by the data is the most intuitive measure of variability. The range is exactly the distance between the smallest data point (min) and the largest one (Max).

Range = Max – min

Note: When we first looked at the histogram, and tried to get a first feel for thespread of the data, we were actually approximating the range, rather than calculating the exact range.

EXAMPLE: Best Actress Oscar Winners

Here we have the Best Actress Oscar winners' data
34 34 26 37 42 41 35 31 41 33 30 74 33 49 38 61 21 41 26 80 43 29 33 35 45
49 39 34 26 25 35 33
In this example:

min = 21 (Marlee Matlin for Children of a Lesser God, 1986) Max = 80 (Jessica Tandy for Driving Miss Daisy, 1989)

The range covered by all the data is 80 – 21 = 59 years.
Variance:
The mean of all squared deviation scores.
Although both the range and its most important spinoff, the interquartile range (discussed in Section 4.7), serve as valid measures of variability, neither is among the statistician's preferred measures of variability. Those roles are reserved for the variance and particularly for its square root, the standard deviation, because these measures serve as key components for other important statistical measures. Accordingly, the variance and standard deviation occupy the same exalted position among measures of variability as does the mean among measures of central tendency. Following the computational procedures described in later sections of this chapter, we could calculate the value of the variance for each of the three distributions in Figure 4.1. Its value equals 0.00 for the least variable distribution, A, 0.29 for the moderately

variable distribution, B, and 3.14 for the most variable distribution, C, in agreement with our intuitive judgments about the relative variability of these three distributions. Reconstructing the Variance To understand the variance better, let's reconstruct it step by step. Although a measure of variability, the variance also qualifies as a type of mean, that is, as the balance point for some distribution. To qualify as a type of mean, the values of all scores must be added and then divided by the total number of scores. In the case of the variance, each original score is re-expressed as a distance or deviation from the mean by subtracting the mean. For each of the three distributions in Figure 4.1, the face values of the seven original scores (shown as numbers along the X axis) have been re-expressed as deviation scores from their mean of 10 (shown as numbers in the boxes). For example, in distribution C, one score coincides with the mean of 10, four scores (two 9s and two 11s) deviate 1 unit from the mean, and two scores (one 7 and one 13) deviate 3 units from the mean, yielding a set of seven deviation scores: one 0, two –1s, two 1s, one –3, and one 3. (Deviation scores above the mean are assigned positive signs; those below the mean are assigned negative signs.) Mean of the Deviations Not a Useful Measure No useful measure of variability can be produced by calculating the mean of these seven deviations, since, as you will recall from Chapter 3, the sum of all deviations from their mean always equals zero. In effect, the sum of all negative deviations always counterbalances the sum of all positive deviations, regardless of the amount of variability in the group.

The standard deviation is to quantify the spread of a distribution by measuring how far the observations are from their mean. The standard deviation gives the average (or typicaldistance) between a data point and the mean.

Standard deviation is the measure of the overall spread (variability) of a data set valuesfrom the mean. The more spread out a data set is, the greater are thedistances from themean and the standard deviation.

There are many notations for the standard deviation: SD, s, Sd, StDev. Here, we'll use SDas an abbreviation for standard deviation, and use s as the symbol.Formula

The sample standard deviation formula is:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Calculation

In order to get a better understanding of the standard deviation, it would beuseful tosee an example of how it is calculated.

EXAMPLE: Video Store Customers

The following are the number of customers who entered a video store in8 consecutivehours: 7, 9, 5, 13, 3, 11, 15, 9

To find the standard deviation of the number of hourly customers:

1. Find the mean, x-bar, of your data:

(7 + 9 + 5 + 13 + 3 + 11 + 15 + 9)/8 = 9

2. Find the deviations from the mean:

- The differences between each observation and the mean here are

  (7 – 9), (9 – 9), (5 – 9), (13 – 9), (3 – 9), (11 – 9), (15 – 9), (9 – 9)
  -2, 0, -4, 4, -6, 2, 6, 0

- Since the standard deviation attempts to measure the average (typical) distance between the data points and their mean, it would make sense to average the deviation we obtained.
- Note, however, that the sum of the deviations is zero.

3. To solve the previous problem, in our calculation, we square each of the deviations.

$(-2)^2$, $(0)^2$, $(-4)^2$, $(4)^2$, $(-6)^2$, $(2)^2$, $(6)^2$, $(0)^2$

4, 0, 16, 16, 36, 4, 36, 0

4. Sum the squared deviations and divide by n – 1:

(4 + 0 + 16 + 16 + 36 + 4 + 36 + 0)/(8 – 1)

(112)/(7) = 16

- This value, the sum of the squared deviations divided by n – 1, is called the variance. However, the variance is not used as a measure of spread directly as the units are the square of the units of the original data.

5. The standard deviation of the data is the square root of the variance calculated in step.

In this case, we have the square root of 16 which is 4. We will use the lower case letter s represent the standard deviation. s = 4

- We take the square root to obtain a measure which is in the original units of the data. The units of the variance of 16 are in “squared customers” which is difficult to interpret.
- The units of the standard deviation are in “customers” which makes this measure ofvariation more useful in practice than the variance.

9. The interpretation of the standard deviation is that on average, the actual number of customers who enter the store each hour is 4 away from 9.

- The standard deviation is the square root of the variance (both population and sample).
- While the sample variance is the positive, unbiased estimator for the population variance, the units for the variance are squared.
- The standard deviation is a common method for numerically describing the distribution of a variable. The population standard deviation is σ (sigma) and sample standard deviation is *s*.

Population standard deviation Sample standard deviation

Example 7

Compute the standard deviation of the sample data: 3, 5, 7 with a sample mean of 5.

DEGREES OF FREEDOM (d f)

Degrees of freedom (df) refers to the number of values that are free to vary, given one or more mathematical restrictions, in a sample being used to estimate a population characteristic.

The number of values free to vary, given one or more mathematical restrictions.

degrees of freedom, that is, df = n – 1.

Inter-Quartile Range (IQR)

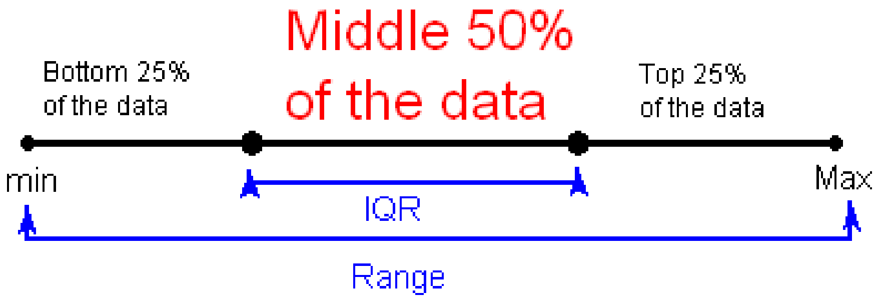
The Inter-Quartile Range or IQR measures the variability of a distribution by giving us the range covered by the MIDDLE 50% of the data.To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.

IQR = Q3 – Q1

Q3 = 3rd Quartile = 75th PercentileQ1 = 1st

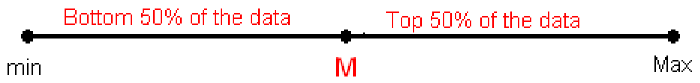
Quartile = 25th Percentile

The following picture illustrates this idea: (Think about the horizontal line as the data ranging from the min to the Max). IMPORTANT NOTE: The "lines" in the following illustrations are not to scale. The equal distances indicate equal amounts of data NOT equal distance between the numeric values.

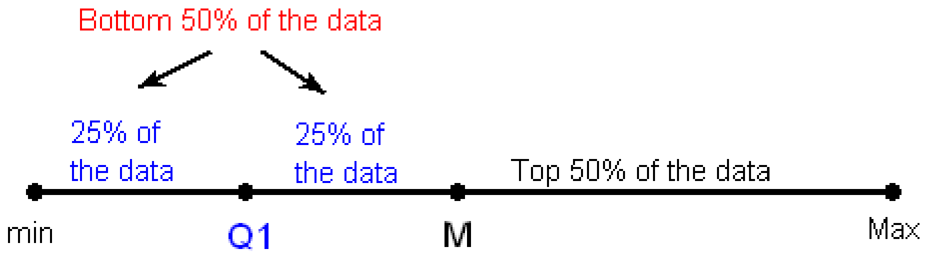


To calculate the IQR:

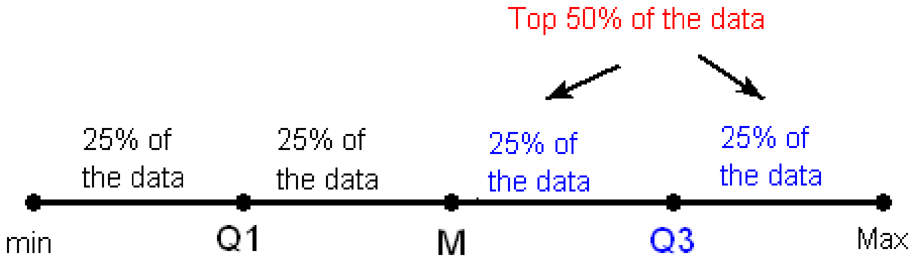
1. Arrange the data in increasing order, and find the median M. Recall that the median divides the data, so that 50% of the data points are below the median, and 50% of the data points are above the median.



2. Find the median of the lower 50% of the data. This is called the first quartile of the distribution, and the point is denoted by Q1. Note from the picture that Q1 divides the lower 50% of the data into two halves, containing 25% of the data points in eachhalf. Q1 is called the first quartile, since one quarter of the data points fall below it.

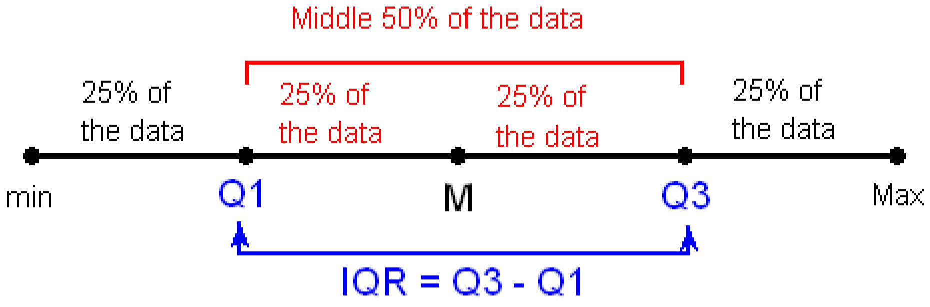


3. Repeat this again for the top 50% of the data. Find the median of the top 50% of the data. This point is called the third quartile of the distribution, and is denoted by Q3.Note from the picture that Q3 divides the top 50% of the data into two halves, with 25%of the data points in each.Q3 is called the third quartile,since three quarters of the data points fall below it.



4. The middle 50% of the data falls between Q1 and Q3, and therefore:

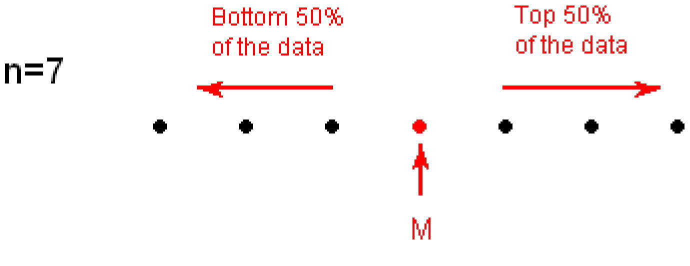
IQR = Q3 – Q1.

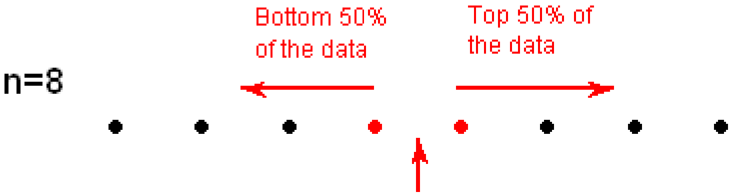


Comments:

1. The last picture shows that Q1, M, and Q3 divide the data into four quarters with 25%of the data points in each, where the median is essentially the second quartile. The use of IQR = Q3 – Q1 as a measure of spread is therefore particularly appropriate when the median M is used as a measure ofcenter.

2. We can define a bit more precisely what is considered the bottom or top 50% of the data. The bottom (top) 50% of the data is all the observations whose position in the ordered list is to the left (right) of the location of the overall median M. The following picture will visually illustrate this for the simple cases of n = 7 and n = 8.

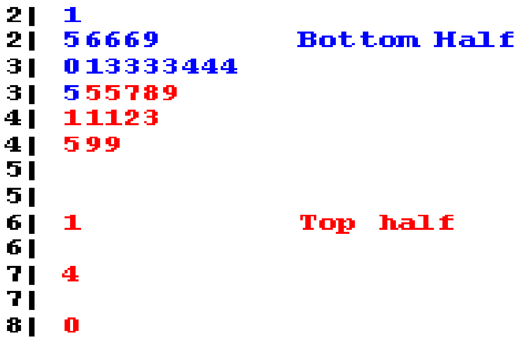




Note that when n is odd (as in n = 7 above), the median is not included in either the bottom or top half of the data; When n is even (as in n = 8 above), the data are naturally divided into two halves.

EXAMPLE: Best Actress Oscar Winners

To find the IQR of the Best Actress Oscar winners' distribution, it will be convenient touse the stemplot.



```
2|  1
2|  56669          Bottom Half
3|  013333444
3|  555789
4|  11123
4|  599
5|
5|
6|  1              Top  half
6|
7|  4
7|
8|  0
```

Q1 is the median of the bottom half of the data. Since there are 16 observations in that half, Q1 is the mean of the 8th and 9th ranked observations in that half:

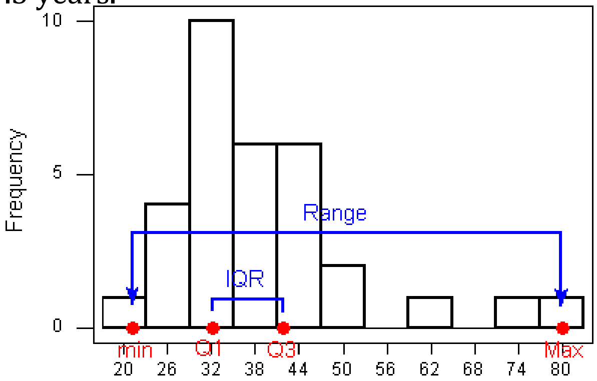
Q1 = (31 + 33) / 2 = 32

Similarly, Q3 is the median of the top half of the data, and since there are 16 observations in that half, Q3 is the mean of the 8th and 9th ranked observations in that half:

Q3 = (41 + 42) / 2 = 41.5

IQR = 41.5 – 32 = 9.5

Note that in this example, the range covered by all the ages is 59 years, while the range covered by the middle 50% of the ages is only 9.5 years. While the whole dataset is spread over a range of 59 years, the middle 50% of the datais packed into only 9.5 years.



The Normal Distribution

Many continuous random variables have a bell-shaped or somewhat symmetric distribution.

This is a normal distribution. In other words, the probability distribution of its relative frequency histogram follows a normal curve.

The curve is bell-shaped, symmetric about the mean, and defined by μ and σ (the mean and standard deviation).

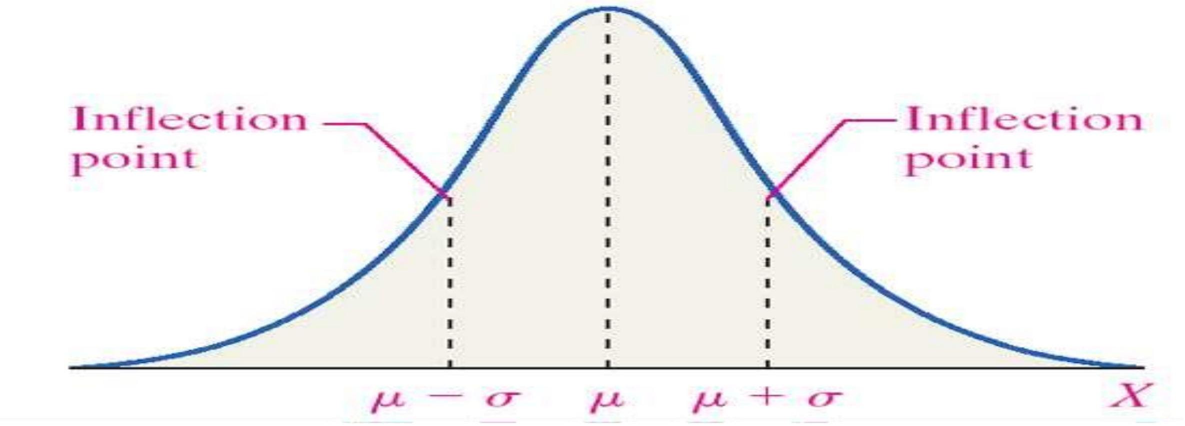


Figure 9. A normal distribution.

There are normal curves for every combination of μ and σ.

- The mean (μ) shifts the curve to the left or right.
- The standard deviation (σ) alters the spread of the curve.
- The first pair of curves have different means but the same standard deviation.
- The second pair of curves share the same mean (μ) but have different standard deviations.
- The pink curve has a smaller standard deviation. It is narrower and taller, and the probability is spread over a smaller range of values.

- The blue curve has a larger standard deviation. The curve is flatter and the tails are

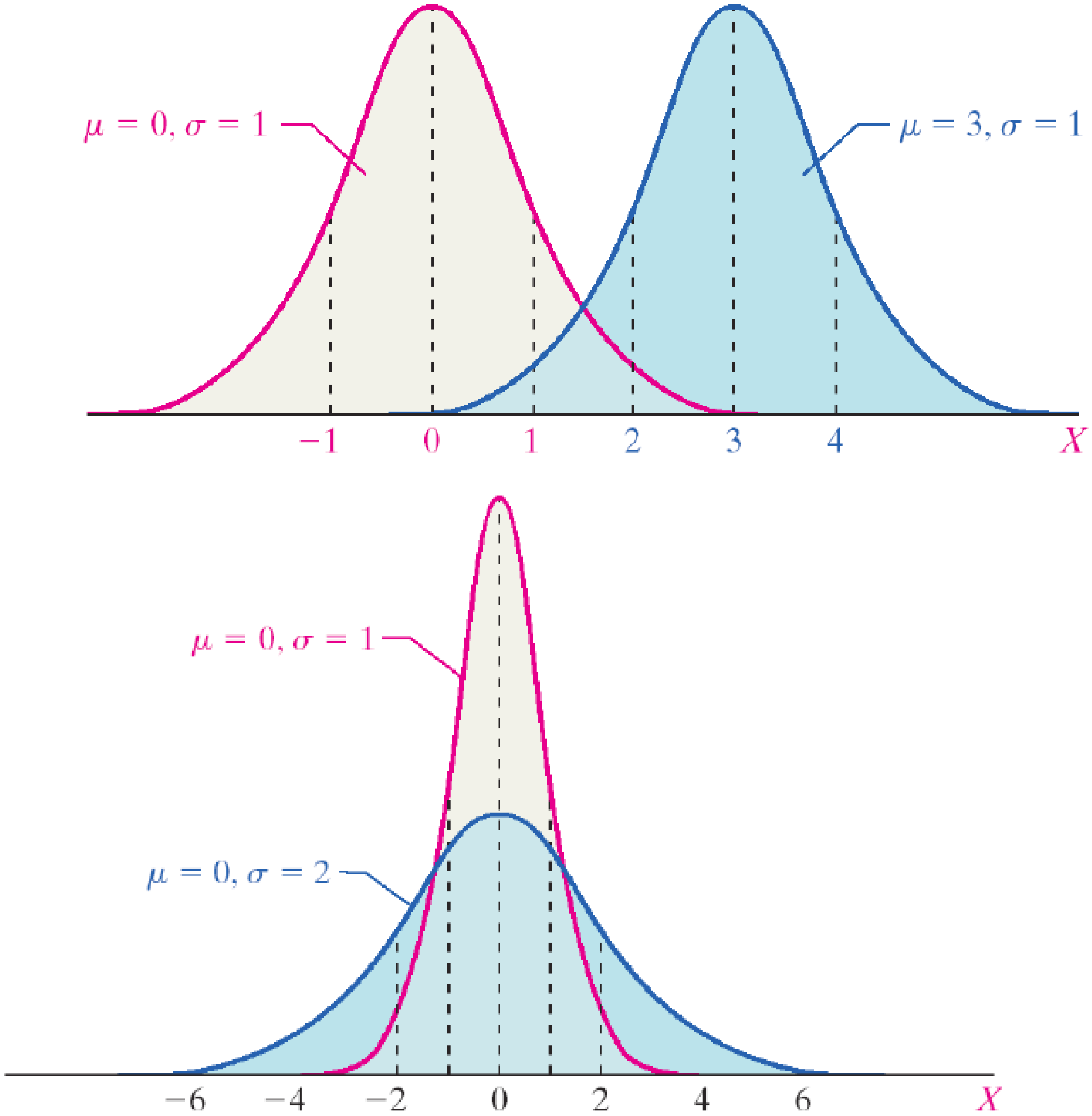


Figure 10. A comparison of normal curves.

Properties of the normal curve:

- The mean is the center of this distribution and the highest point.
- The curve is symmetric about the mean. (The area to the left of the mean equals the area to the right of the mean.)
- The total area under the curve is equal to one.
- As *x* increases and decreases, the curve goes to zero but never touches.

- The PDF of a normal curve is $y = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$.
- A normal curve can be used to estimate probabilities.
- A normal curve can be used to estimate proportions of a population that have certain x-values.

The Standard Normal Distribution

There are millions of possible combinations of means and standard deviations for continuous random variables.

Finding probabilities associated with these variables would require us to integrate the PDF over the range of values we are interested in.

To avoid this, we can rely on the standard normal distribution. T

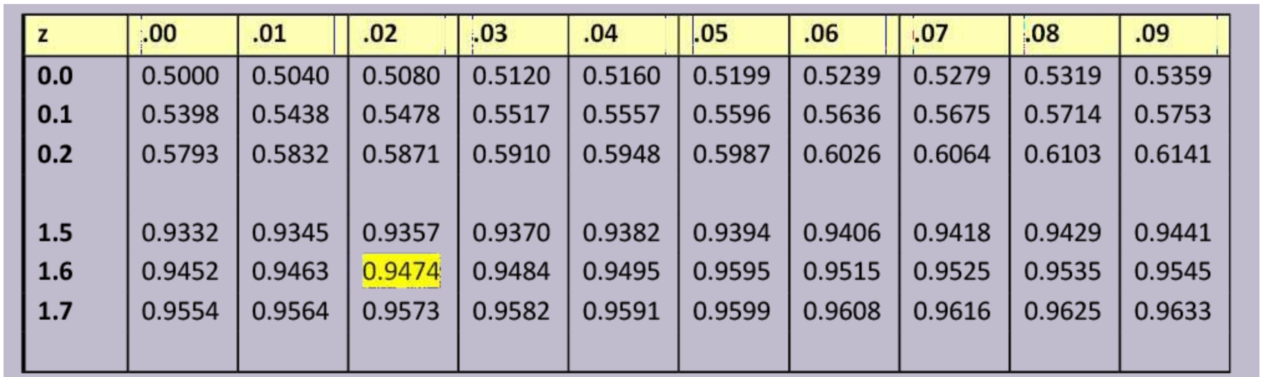
he standard normal distribution is a special normal distribution with a μ = 0 and **σ** = 1. We can use the Z-score to standardize any normal random variable, converting the x-values to Z-scores, thus allowing us to use probabilities from the standard normal table. So how do we find area under the curve associated with a Z-score?

Standard Normal Table

- The standard normal table gives probabilities associated with specific Z-scores.
- The table we use is cumulative from the left.
- The negative side is for all Z-scores less than zero (all values less than the mean).
- The positive side is for all Z-scores greater than zero (all values greater than the mean).
- Not all standard normal tables work the same way.

Example 10

What is the area associated with the Z-score 1.62?



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9595	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

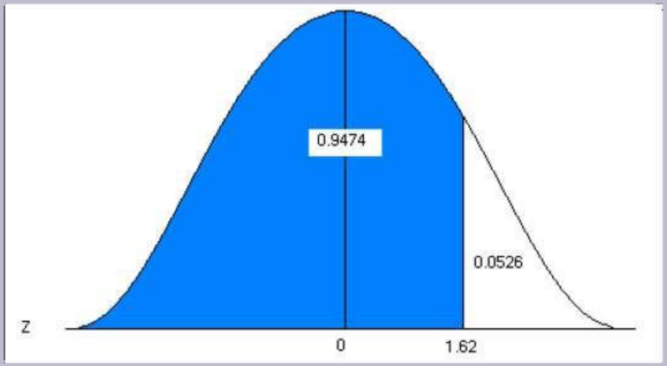


Figure 11. The standard normal table and associated area for z = 1.62.

Reading the Standard Normal Table

- Read down the Z-column to get the first part of the Z-score (1.6).
- Read across the top row to get the second decimal place in the Z-score (0.02).
- The intersection of this row and column gives the area under the curve to the left of the Z-score.

Finding Z-scores for a Given Area

- What if we have an area and we want to find the Z-score associated with that area?
- Instead of Z-score → area, we want area → Z-score.
- We can use the standard normal table to find the area in the body of values and read backwards to find the associated Z-score.
- Using the table, search the probabilities to find an area that is closest to the probability you are interested in.

Example 11

To find a Z-score for which the area to the right is 5%:

Since the table is cumulative from the left, you must use the complement of 5%.

1.000 – 0.05 = 0.9500

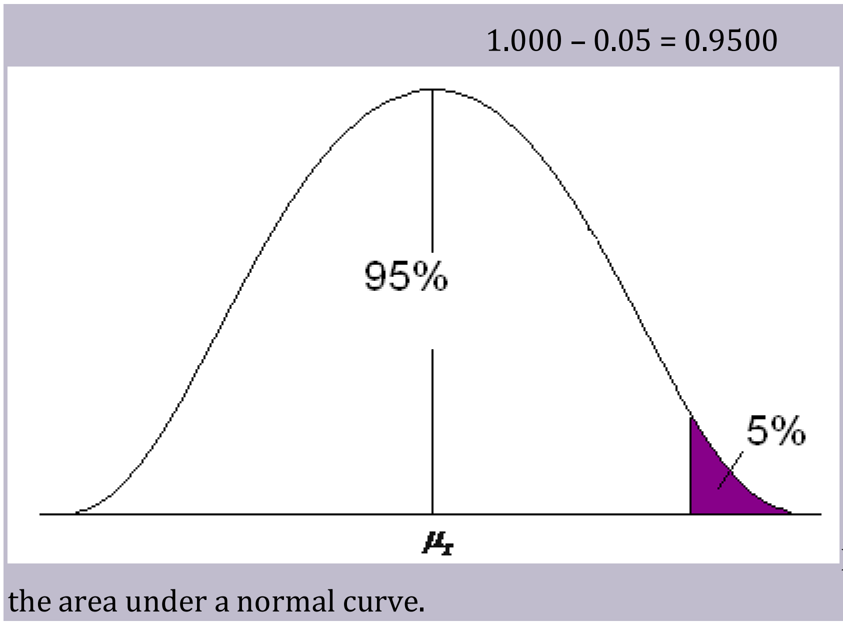


Figure 12. The upper 5% of the area under a normal curve.

- Find the Z-score for the area of 0.9500.
- Look at the probabilities and find a value as close to 0.9500 as possible.

Figure 13. The standard normal table.

The Z-score for the 95th percentile is 1.64.**Area in between Two Z-scores**

Example 12

To find Z-scores that limit the middle 95%:

- The middle 95% has 2.5% on the right and 2.5% on the left.
- Use the symmetry of the curve.

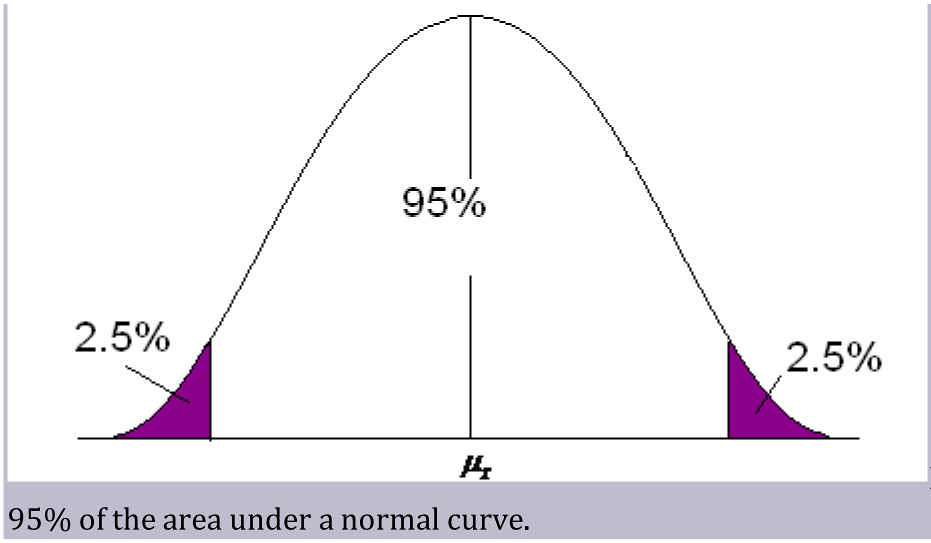


Figure 14. The middle 95% of the area under a normal curve.

- Look at your standard normal table. Since the table is cumulative from the left, it is easier to find the area to the left first.
- Find the area of 0.025 on the negative side of the table.
- The Z-score for the area to the left is -1.96.
- Since the curve is symmetric, the Z-score for the area to the right is 1.96.

Common Z-scores

There are many commonly used Z-scores:

- Z.05 = 1.645 and the area between -1.645 and 1.645 is 90%
- Z.025 = 1.96 and the area between -1.96 and 1.96 is 95%
- Z.005 = 2.575 and the area between -2.575 and 2.575 is 99%

Applications of the Normal Distribution

Typically, our normally distributed data do not have $\mu = 0$ and $\boldsymbol{\sigma} = 1$, but we can relate any normal distribution to the standard normal distributions using the Z-score. We can transform values of x to values of z.

$$z = \frac{x - \mu}{\sigma}$$

For example, if a normally distributed random variable has a $\mu = 6$ and $\boldsymbol{\sigma} = 2$, then a value of $x = 7$ corresponds to a Z-score of 0.5.

$$Z = \frac{7-6}{2} = 0.5$$

This tells you that 7 is one-half a standard deviation above its mean. We can use this relationship to find probabilities for any normal random variable.

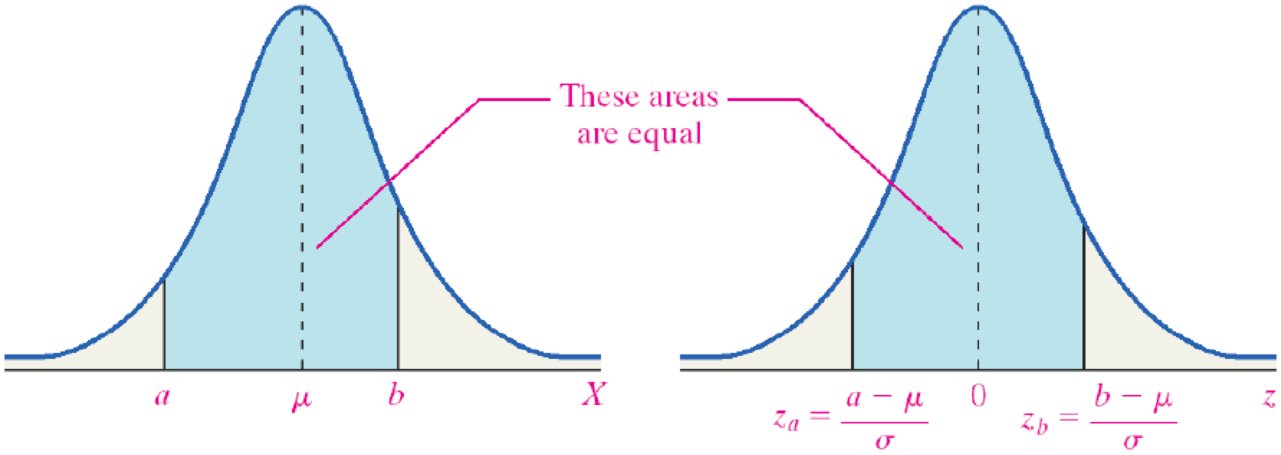


Figure 15. A normal and standard normal curve.

To find the area for values of X, a normal random variable, draw a picture of the area of interest, convert the x-values to Z-scores using the Z-score and then use the standard normal table to find areas to the left, to the right, or in between.

$$z = \frac{x-\mu}{\sigma}$$

Example 13

Adult deer population weights are normally distributed with µ = 110 lb. and **σ** = 29.7 lb. As a biologist you determine that a weight less than 82 lb. is unhealthy and you want to know what proportion of your population is unhealthy.

P(x<82)

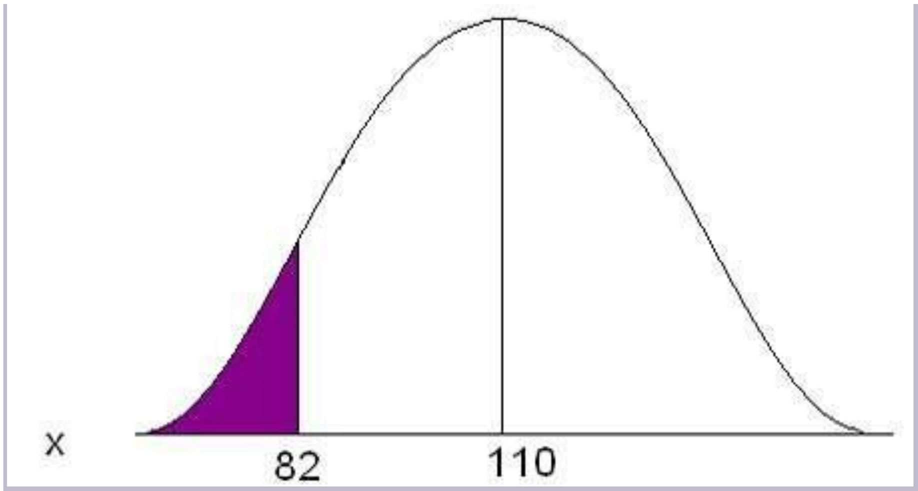


Figure 16. The area under a normal curve for P(x<82).

Convert 82 to a Z-score $$z = \frac{82 - 110}{29.7} = -0.94$$

The *x* value of 82 is 0.94 standard deviations below the mean.

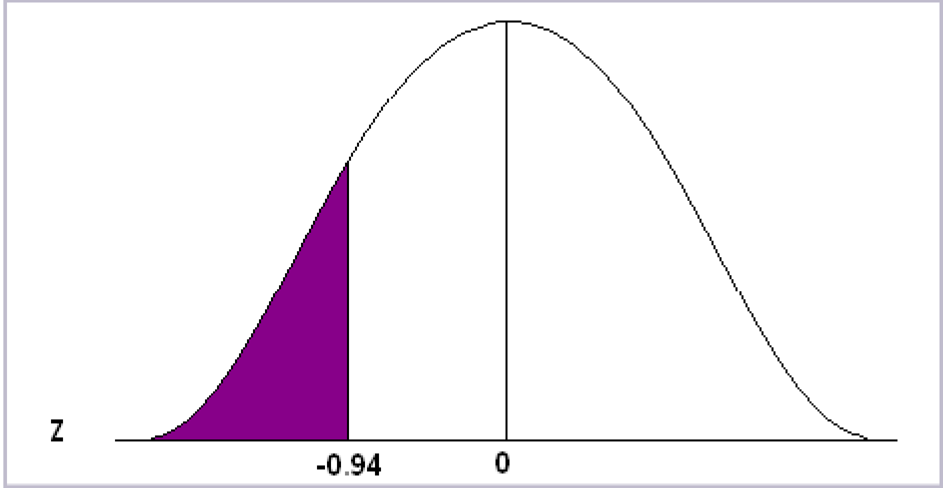


Figure 17. Area under a standard normal curve for P(z<-0.94).

Go to the standard normal table (negative side) and find the area associated with a Z-score of -0.94.

This is an "area to the left" problem so you can read directly from the table to get the probability.

P(x<82) = 0.1736

Approximately 17.36% of the population of adult deer is underweight, OR one deer chosen at random will have a 17.36% chance of weighing less than 82 lb.

Example 14

Statistics from the Midwest Regional Climate Center indicate that Jones City, which has a large wildlife refuge, gets an average of 36.7 in. of rain each year with a standard deviation of 5.1 in. The amount of rain is normally distributed. During what percent of the years does Jones City get more than 40 in. of rain?

P(x > 40)

Figure 18. Area under a normal curve for P(x>40).

$$P(x>40) = (1-0.7422) = 0.2578$$

For approximately 25.78% of the years, Jones City will get more than 40 in. of rain.

Assessing Normality

- If the distribution is unknown and the sample size is not greater than 30 (Central Limit Theorem), we have to assess the assumption of normality.
- Our primary method is the normal probability plot. This plot graphs the observed data, ranked in ascending order, against the "expected" Z-score of that rank.
- If the sample data were taken from a normally distributed random variable, then the plot would be approximately linear.
- Examine the following probability plot.
- The center line is the relationship we would expect to see if the data were drawn from a perfectly normal distribution.

- Notice how the observed data (red dots) loosely follow this linear relationship. Minitab also computes an Anderson-Darling test to assess normality.

- The null hypothesis for this test is that the sample data have been drawn from a normally distributed population. A p-value greater than 0.05 supports the assumption of normality.

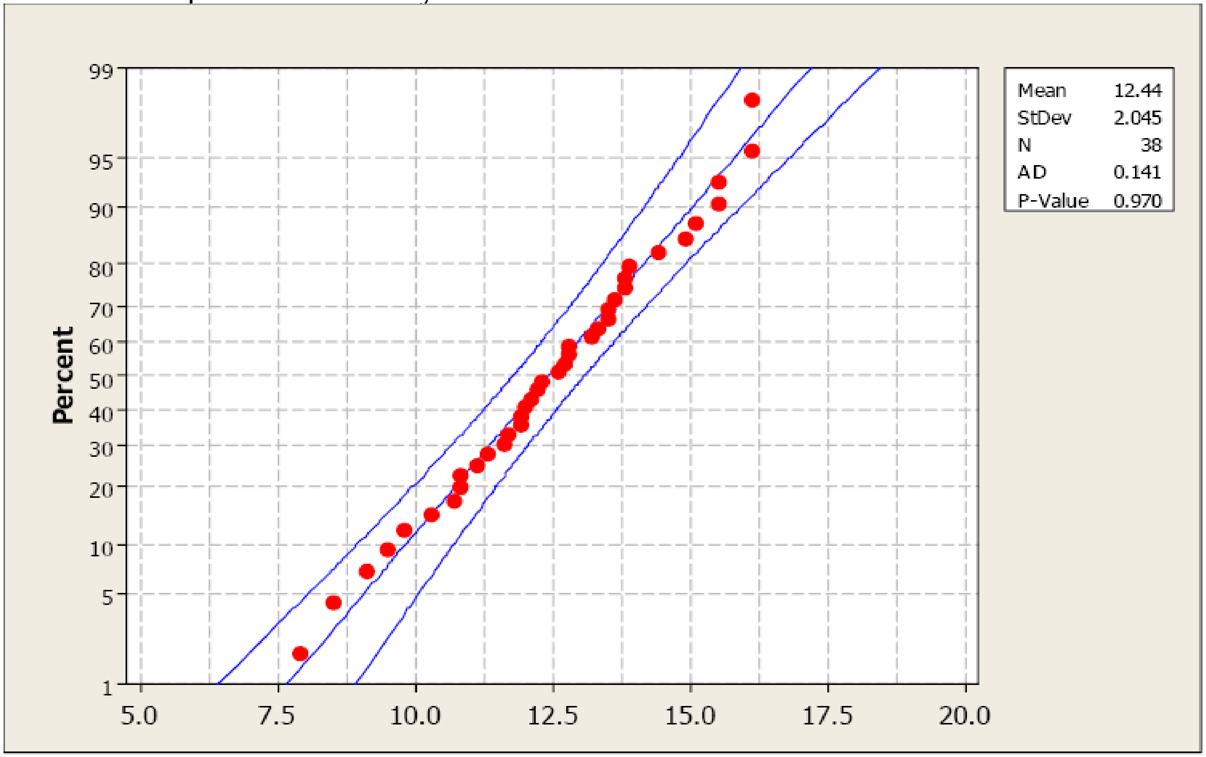


Figure 19. A normal probability plot generated using Minitab 16.

Compare the histogram and the normal probability plot in this next example. The histogram indicates a skewed right distribution.

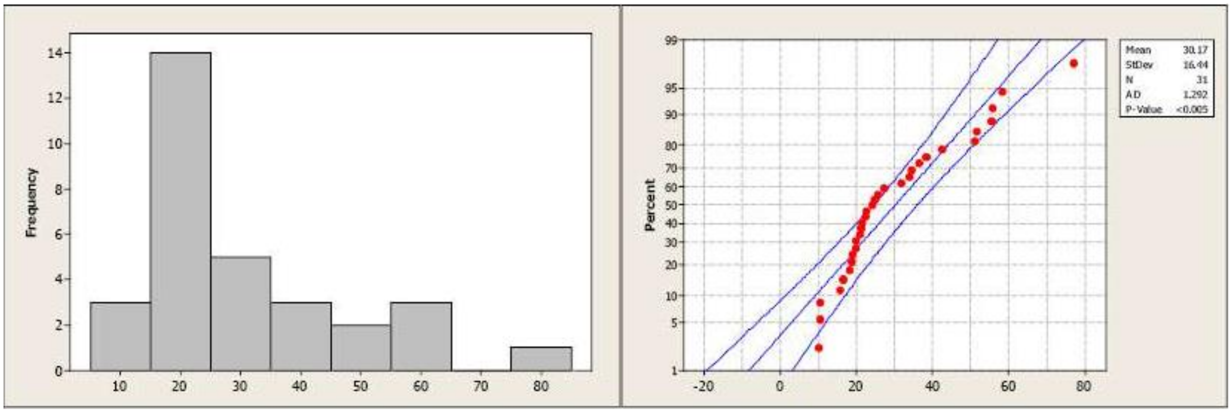


Figure 20. Histogram and normal probability plot for skewed right data.

The observed data do not follow a linear pattern and the p-value for the A-D test is less than 0.005 indicating a non-normal population distribution.

Normality cannot be assumed. You must always verify this assumption. Remember, the probabilities we are finding come from the standard NORMAL table. If our data are NOT normally distributed, then these probabilities DO NOT APPLY.

UNIT III DESCRIBING RELATIONSHIPS

Correlation –Scatter plots –correlation coefficient for quantitative data –computational formula for correlation coefficient – Regression –regression line –least squares regression line – Standard error of estimate – interpretation of r2 –multiple regression equations – regression towards the mean

3.1 Scatterplot

- ❖ Define scatter plot(2M)
- ❖ Explain scatter plot with example(16M)
- ❖ How to interpret scatter plots(16M)

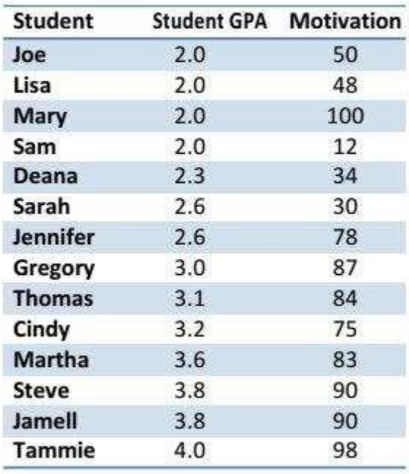
- The most useful graph for displaying the relationship between two quantitative variables is a **scatterplot**.

> A **scatterplot** shows the relationship between two quantitative variables measured for the same individuals. The values of one variable appear on the horizontal axis, and the values of the othervariable appear on the vertical axis. Each individual in the data appears as a point on the graph.

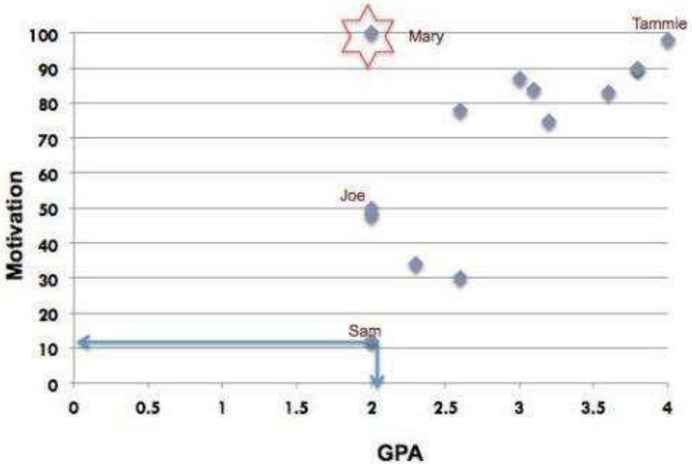
- Many research projects are **correlational studies** because they investigate the relationships that may exist between variables. Prior to investigating the relationship between two quantitative variables, it is always helpful to create a graphical representation that includes both of these variables. Such a graphical representation is called a **scatterplot**.

3.1.1 Scatterplot

What is the relationship between students' achievement motivation and GPA?



Student	Student GPA	Motivation
Joe	2.0	50
Lisa	2.0	48
Mary	2.0	100
Sam	2.0	12
Deana	2.3	34
Sarah	2.6	30
Jennifer	2.6	78
Gregory	3.0	87
Thomas	3.1	84
Cindy	3.2	75
Martha	3.6	83
Steve	3.8	90
Jamell	3.8	90
Tammie	4.0	98



- In this example, the relationship between students' achievement motivation and their GPA is being investigated.
- The table on the left includes a small group of individuals for whom GPA and scores on a motivation scale have been recorded. GPAs can range from 0 to 4 and motivation scores in this example range from 0 to 100. Individuals in this table were ordered based on their GPA.
- Simply looking at the table shows that, in general, as GPA increases, motivation scores also increase.
- However, with a real set of data, which may have hundreds or even thousands of individuals, a pattern cannot be detected by simply looking atthe numbers. Therefore, a very useful strategy is to represent the two variables graphically to illustrate the relationship between them.
- A graphical representation of individual scores on two variables is called a **scatterplot**.
- The image on the right is an example of a scatterplot and displays the datafrom the table on the left. GPA scores are displayed on the horizontal axisand motivation scores are displayed on the vertical axis.
- Each dot on the scatterplot represents one individual from the data set. The location of each point on the graph depends on both the GPA and motivation scores. Individuals with higher GPAs are located further to the right and individuals with higher motivation scores are located higher up on the graph.
- Sam, for example, has a GPA of 2 so his point is located at 2 on the right. He also has a motivation score of 12, so his point is located at 12 going up.
- Scatterplots are not meant to be used in great detail because there areusually hundreds of individuals in a data set.
- The purpose of a scatterplot is to provide a general illustration of the relationship between the two variables.
- In this example, in general, as GPA increases so does an individual'smotivation score.
- One of the students in this example does not seem to follow the general pattern: Mary. She is one of the students with the lowest GPA, but she has the maximum score on the motivation scale. This makes her an exception oran outlier.

3.1.2 Interpreting Scatterplots

How to Examine a Scatterplot

As in any graph of data, look for the *overall pattern* and for striking *departures* from that pattern.

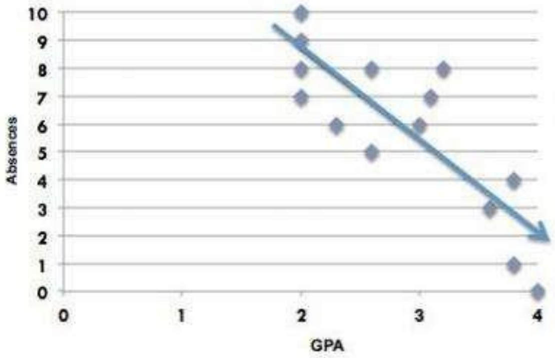
- The overall pattern of a scatterplot can be described by the **direction**, **form**, and **strength** of the relationship.
- An important kind of departure is an **outlier**, an individual value that falls outside the overall pattern of the relationship.

Interpreting Scatterplots: Direction

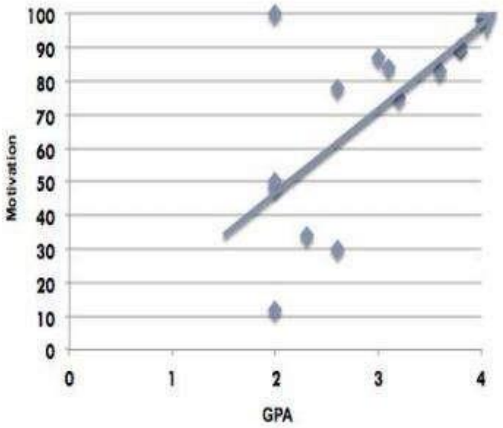
- One important component to a scatterplot is the **direction** of the relationshipbetween the two variables.

Two variables have a **positive association** when above-average values of one tend to accompany above-average values of the other, and when below-average values also tend to occur together.

Two variables have a **negative association** when above-average values of one tend to accompany below-average values of the other.



This example comparesstudents' achievement
motivation and their GPA. These two variables have a **positive association** becauseas GPA increases, so does motivation.

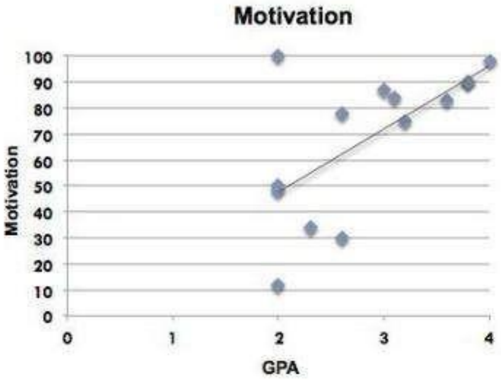


This example compares students' GPA and their numberof absences. These two variables have a **negative association** because, in general,as a student's number of absences decreases, their GPA increases

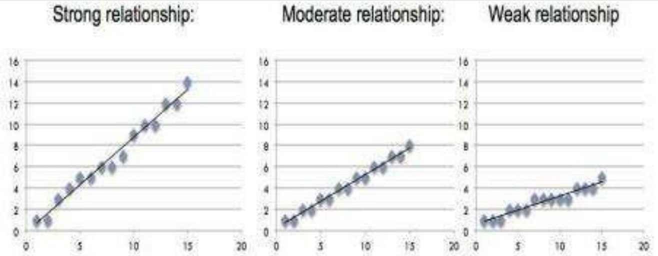
Interpreting Scatterplots: Form

- Another important component to a scatterplot is the **form** of the relationshipbetween the two variables.

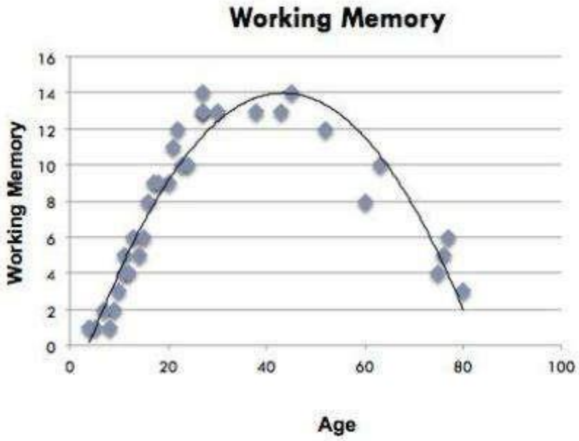
Linear relationship:



This example illustrates a linear relationship. This means that the points on the scatterplot closely resemble a straight line. A relationship is linear if one variable increases by approximately the same rate as theother variables changes by one unit.


Strong relationship: Moderate relationship: Weak relationship

Curvilinear relationship:



This example illustrates a relationship that has the form of acurve, rather than a straight line. This is due to the fact that one variable does not increase at a constant rate and may even start decreasing after a certain point.

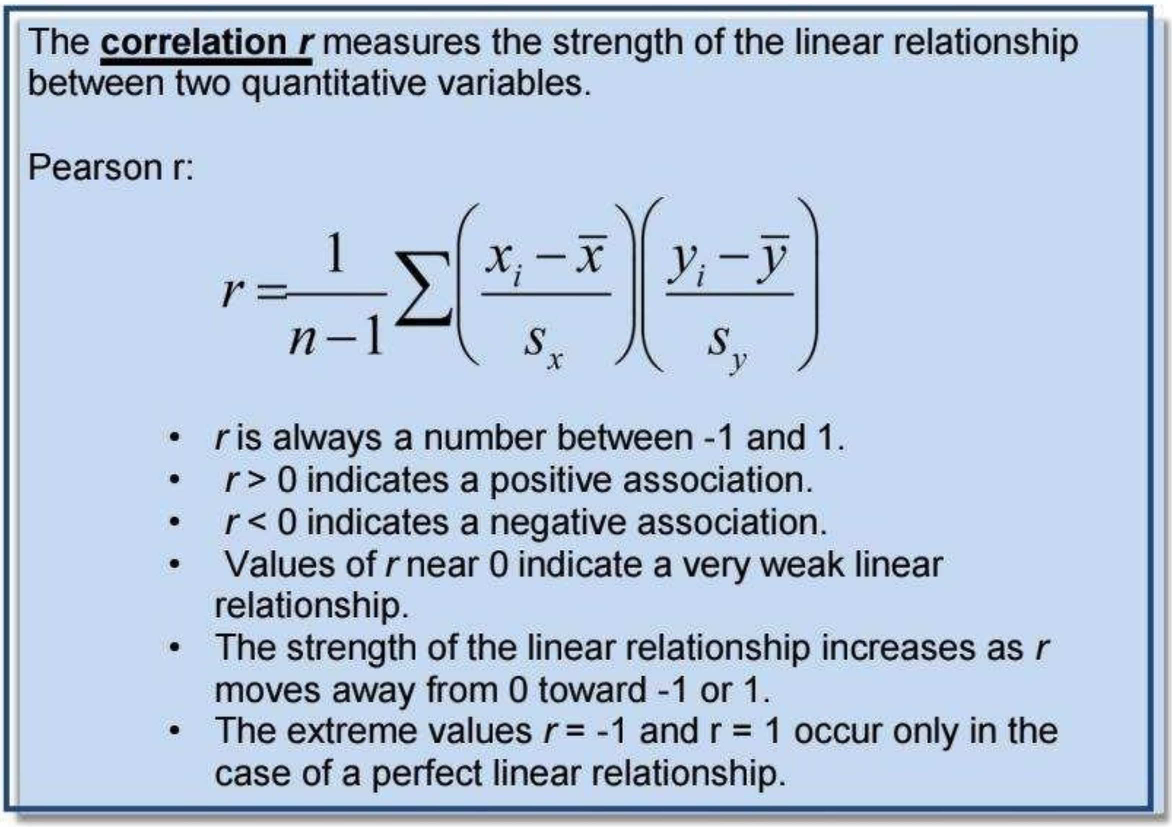
This example describes a curvilinear relationship between the variable "age" and the variable "working memory." In this example, working memory increases throughout childhood, remains steady in adulthood, and begins decreasing around age 50.

Interpreting Scatterplots: Strength

- Another important component to a scatterplot is the **strength** of therelationship between the two variables.
- The **slope** provides information on the strength of the relationship.
- The strongest linear relationship occurs when the slope is 1. This means thatwhen one variable increases by one, the other variable also increases by the same amount. This line is at a 45 degree angle.
- The strength of the relationship between two variables is a crucial piece of information. Relying on the interpretation of a scatterplot is too subjective. More precise evidence is needed, and this evidence is obtained by computinga coefficient that measures the strength of the relationship under investigation.

Measuring Linear Association

- A scatterplot *displays* the strength, direction, and form of the relationshipbetween two quantitative variables.
- A correlation coefficient *measures* the strength of that relationship.



The **correlation *r*** measures the strength of the linear relationship between two quantitative variables.

Pearson r:

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- *r* is always a number between -1 and 1.
- *r* > 0 indicates a positive association.
- *r* < 0 indicates a negative association.
- Values of *r* near 0 indicate a very weak linear relationship.
- The strength of the linear relationship increases as *r* moves away from 0 toward -1 or 1.
- The extreme values *r* = -1 and r = 1 occur only in the case of a perfect linear relationship.

- Calculating a Pearson correlation coefficient requires the assumption that therelationship between the two variables is linear.
- There is a rule of thumb for interpreting the strength of a relationship basedon its r value (use the absolute value of the r value to make all values positive):

Absolute Value of r	Strength of Relationship
r < 0.3	None or very weak
0.3 < r <0.5	Weak
0.5 < r < 0.7	Moderate
r > 0.7	Strong

- The relationship between two variables is generally considered strong whentheir r value is larger than 0.7.

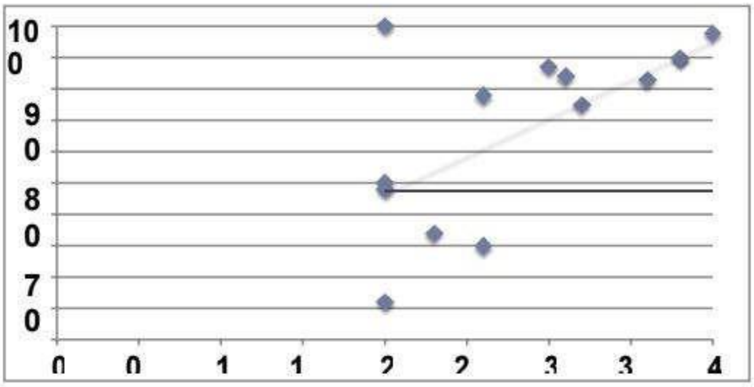
3.2 Correlations

- **What is correlations?(2M)**
- **Facts about correlations(2M)**

Example: There is a moderate, positive, linear relationship between GPA andachievement motivation.

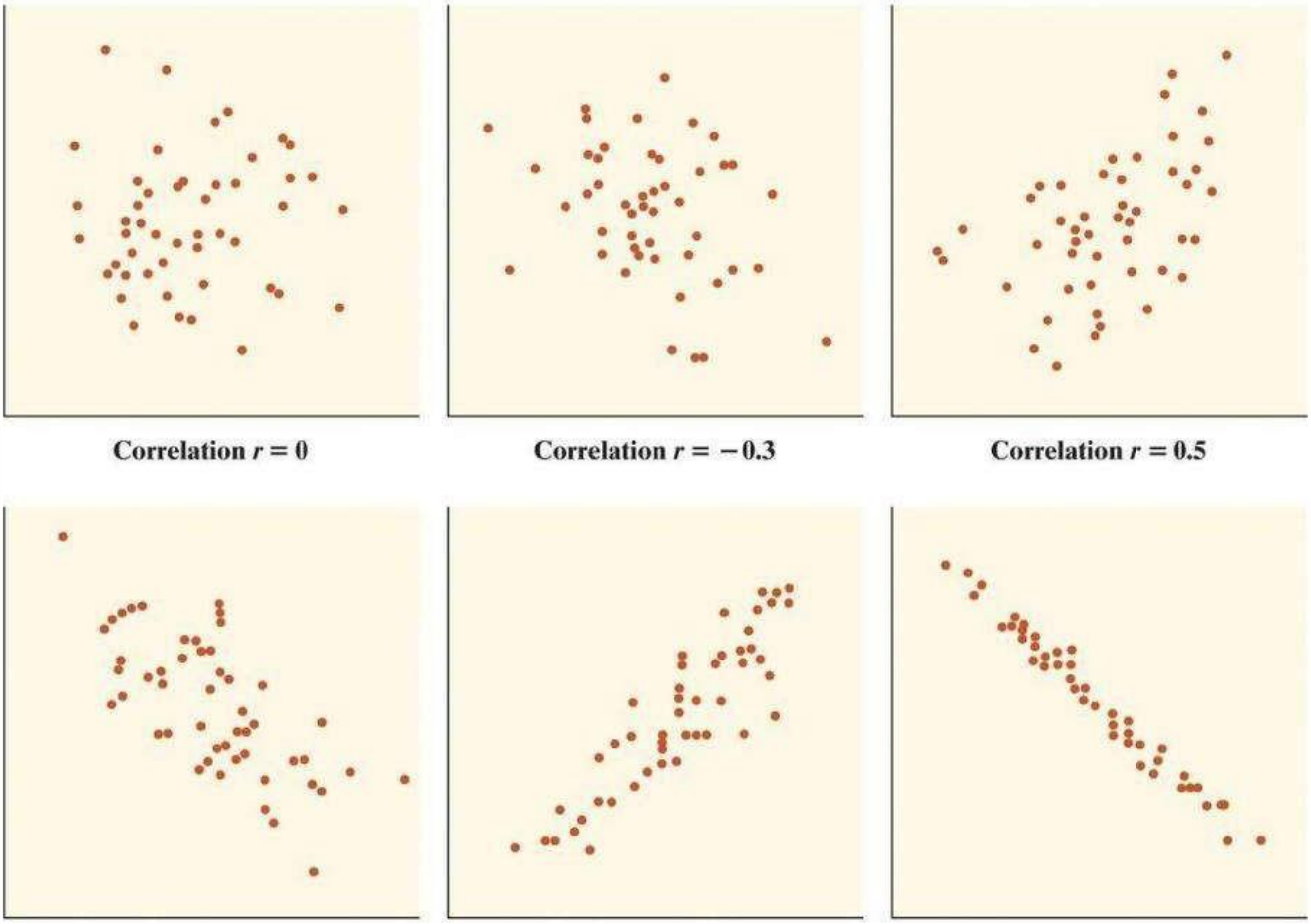
- Based on the criteria listed on the previous page, the value of r in this case (r = 0.62) indicates that there is a positive, linear relationship of **moderate** strength between achievement motivation and GPA.

r = 0.62



Correlation

- The images below illustrate what the relationships might look like atdifferent degrees of strength (for different values of r).



Correlation $r = 0$

Correlation $r = -0.3$

Correlation $r = 0.5$

Correlation $r = -0.7$

Correlation $r = 0.9$

Correlation $r = -0.99$

- For a correlation coefficient of zero, the points have no direction, the shape isalmost round, and a line does not fit to the points on the graph.
- As the correlation coefficient increases, the observations group closertogether in a linear shape.
- The line is difficult to detect when the relationship is weak (e.g., r = -0.3), butbecomes more clear as relationships become stronger (e.g., r = -0.99)

Facts About Correlation

1) The order of variables in a correlation is not important.
2) Correlations provide evidence of association, not causation.
3) *r* has no units and does not change when the units of measure of *x*, *y*, or bothare changed.
4) Positive *r* values indicate positive association between the variables, andnegative *r* values indicate negative associations.
5) The correlation *r* is always a number between -1 and 1.

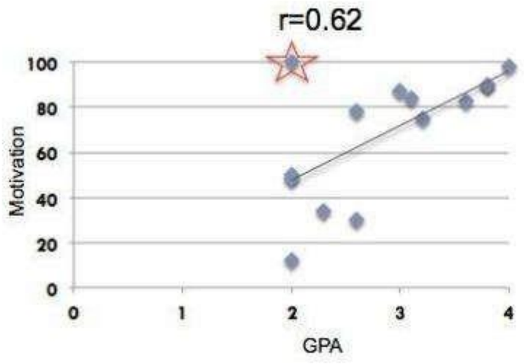
Pearson *r*: Assumptions

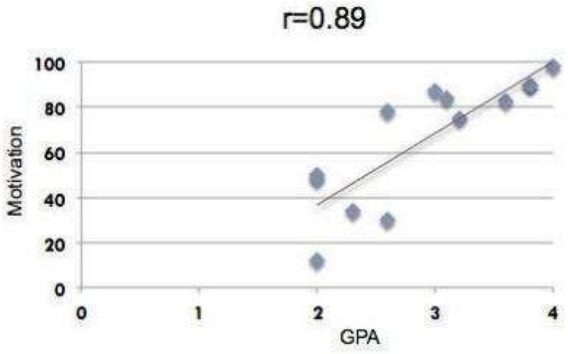
Assumptions:

- Correlation requires that both variables be quantitative.
- Correlation describes *linear* relationships. Correlation does not describe curve relationships between variables, no matter how strong the relationshipis.

Cautions:

- Correlation is not resistant. *r* is strongly affected by outliers.
- Correlation is not a complete summary of two-variable data.
- For example:



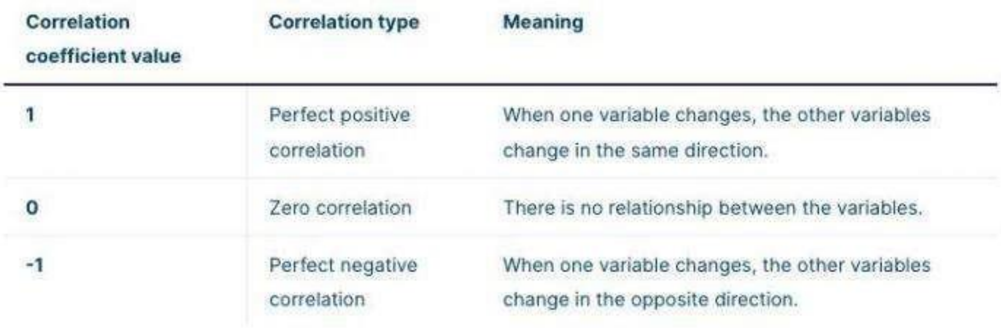


- The correlation coefficient is based on means and standard deviations, so it isnot robust to outliers; it is strongly affected by extreme observations. These individuals are sometimes referred to as *influential observations* because they have a strong impact on the correlation coefficient.
- For instance, in the above example the correlation coefficient is 0.62 on the left when the outlier is included in the analysis. However, when this outlier is removed, the correlation coefficient increases significantly to 0.89.
- This one case, when included in the analysis, reduces a strong relationship toa moderate relationship.
- This case makes such a big difference in this example because the data set contains a very small number of individuals. As a general rule, as the size ofthe sample increases, the influence of extreme observations decreases.
- When describing the relationship between two variables, correlations are just one piece of the puzzle. This information is necessary, but not sufficient.Other analyses should also be conducted to provide more information.

CORRELATION COEFFICIENT:

- What does a correlation coefficient tell you?(2M)
- Significance of correlation coefficient(2M)
- How to interpret correlation coefficient?(8M)
- Types of Correlation coefficient(16M)
- What are the assumptions our data has to meet for pearson'sr?(2M)
- Give Pearson's r formula with explaination(2M)
- Give spearmans's rho formula(2M)

A correlation coefficient is a number between -1 and 1 that tells you the strength and direction of a relationship between variables. In other words, it reflects how similar the measurements of two or more variables are across a dataset.



Correlation coefficient value	Correlation type	Meaning
1	Perfect positive correlation	When one variable changes, the other variables change in the same direction.
0	Zero correlation	There is no relationship between the variables.
-1	Perfect negative correlation	When one variable changes, the other variables change in the opposite direction.

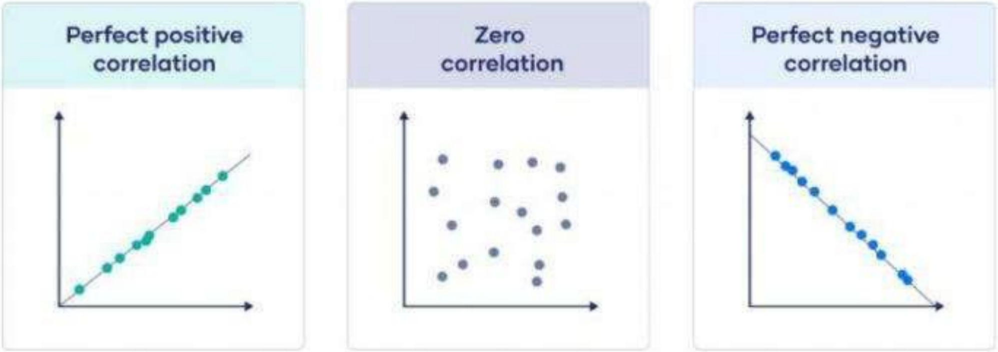


Figure : Co-relation

3.3 Correlation Coefficients

<u>The Statistical Significance of Correlation Coefficients</u>:

- Correlation coefficients have a probability (p-value), which shows **the probability that the relationship between the two variables is equal tozero** (null hypotheses; no relationship).
- **Strong** correlations have **low** p-values because the probability that they haveno relationship is very low.
- Correlations are typically considered statistically significant if the p-value islower than 0.05 in the social sciences, but the researcher has the liberty to decide the p-value for which he or she will consider the relationship to be significant.
- The value of p for which a correlation will be considered statisticallysignificant is called the **<u>alpha level</u>** and must be reported.
- SPSS notation for p values: Sig. (2 tailed)

In the previous example, r = 0.62 and p-value = 0.03. The p-value of 0.03 is less thanthe acceptable alpha level of 0.05, meaning the correlation is statistically significant.

Four things must be reported to describe a relationship:

1) The **strength** of the relationship given by the correlation coefficient.
2) The **direction** of the relationship, which can be positive or negative based onthe sign of the

correlation coefficient.

3) The **shape** of the relationship, which must always be linear to computer aPearson correlation coefficient.
4) Whether or not the relationship is **statistically significant**, which is basedon the p-value

What does a correlation coefficient tell you?

Correlation coefficients summarize data and help you compare results between studies.

Summarizing data

A correlation coefficient is a descriptive statistic. That means that it summarizes sample data without letting you infer anything about the population. A correlation coefficient is a bivariate statistic when it summarizes the relationship between two variables, and it's a multivariate statistic when you have more than two variables.

If your correlation coefficient is based on sample data, you'll need an inferential statistic if you want to generalize your results to the population. You can use an F test or a t test to calculate a test statistic that tells you the statistical significance of your finding.

Comparing studies

A correlation coefficient is also an effect size measure, which tells you the practical significance of a result. Correlation coefficients are unit-free, which makes it possible to directly compare coefficients between studies.

Using a correlation coefficient

In correlational research, you investigate whether changes in one variable are associated with changes in other variables.

Correlational research example

You investigate whether standardized scores from high school are related to academic grades in college. You predict that there's a positive correlation: higher SAT scores are associated with higher college GPAs while lower SAT scores are associated with lower college GPAs.

After data collection, you can visualize your data with a scatterplot by plotting one variable on the x-axis and the other on the y-axis. It doesn't matter which variable you place on either axis.

Visually inspect your plot for a pattern and decide whether there is a linear or non-linear pattern between variables. A linear pattern means you can fit a straight line of best fit between the data points, while a non-linear or curvilinear pattern can take all sorts of different shapes, such as a U-shape or a line with a curve.

Visual inspection example

You gather a sample of 5,000 college graduates and survey them on their high school SAT scores and college GPAs. You visualize the data in a scatterplot to check for a linear

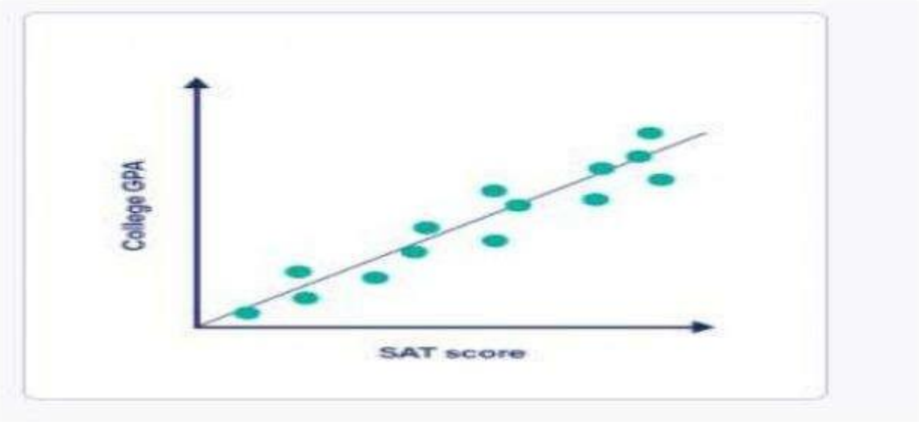


Figure :

There are many different correlation coefficients that you can calculate. After removing any outliers, select a correlation coefficient that's appropriate based on the general shape of the scatter plot pattern. Then you can perform a correlation analysis to find the correlation coefficient for your data.

You calculate a correlation coefficient to summarize the relationship between variables without drawing any conclusions about causation.

Correlation analysis example

You check whether the data meet all of the assumptions for the Pearson's r correlation test.
Both variables are quantitative and normally distributed with no outliers, so you calculate a Pearson's r correlation coefficient.

The correlation coefficient is strong at .58

Interpreting a correlation coefficient

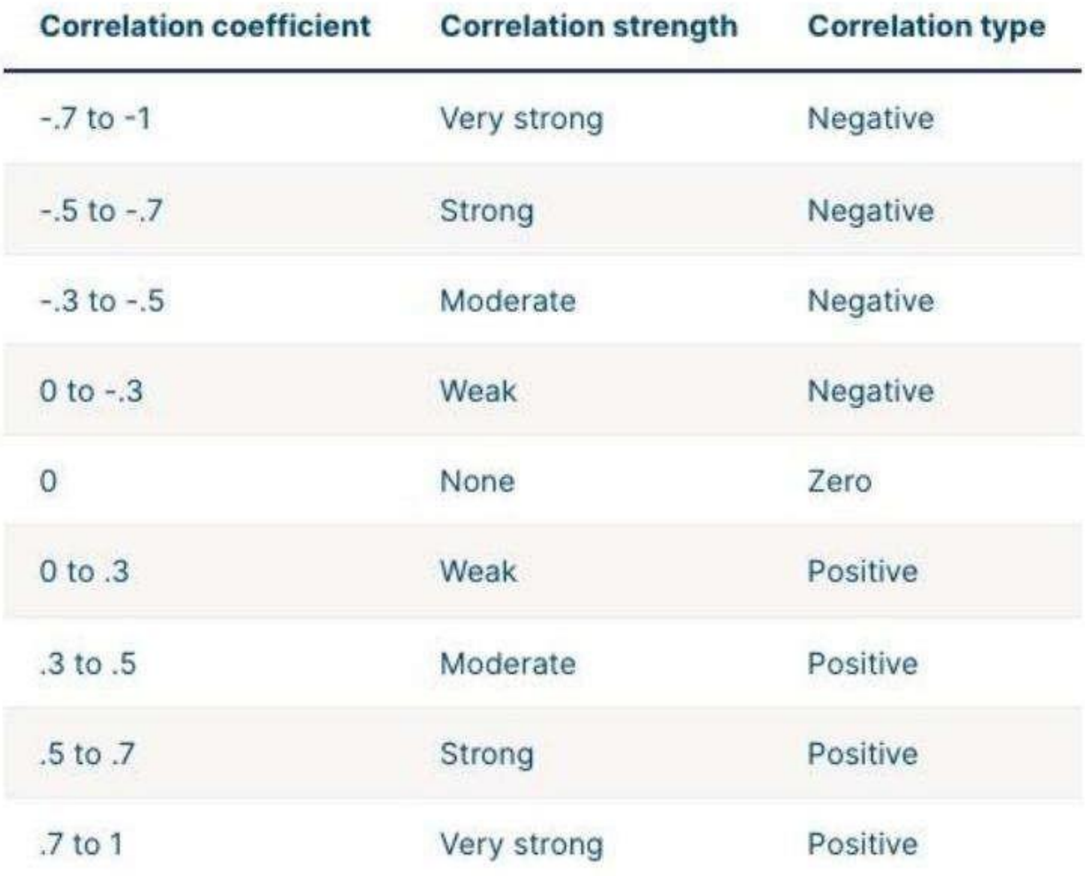
The value of the correlation coefficient always ranges between 1 and -1, and you treat it as a general indicator of the strength of the relationship between variables.

The sign of the coefficient reflects whether the variables change in the same or opposite directions: a positive value means the variables change together in the same direction, while a negative value means they change together in opposite directions.

The absolute value of a number is equal to the number without its sign. The absolute value of a correlation coefficient tells you the magnitude of the correlation: the greater the absolute value, the stronger the correlation.

There are many different guidelines for interpreting the correlation coefficient because findings can vary a lot between study fields. You can use the table below as a general guideline for interpreting correlation strength from the value of the correlation coefficient.

While this guideline is helpful in a pinch, it's much more important to take your research context and purpose into account when forming conclusions. For example, if most studies in your field have correlation coefficients nearing .9, a correlation coefficient of .58 may be low in that context.



Correlation coefficient	Correlation strength	Correlation type
-.7 to -1	Very strong	Negative
-.5 to -.7	Strong	Negative
-.3 to -.5	Moderate	Negative
0 to -.3	Weak	Negative
0	None	Zero
0 to .3	Weak	Positive
.3 to .5	Moderate	Positive
.5 to .7	Strong	Positive
.7 to 1	Very strong	Positive

Table :

Visualizing linear correlations

The correlation coefficient tells you how closely your data fit on a line. If you have a linear relationship, you'll draw a straight line of best fit that takes all of your data points into account on a scatter plot.

The closer your points are to this line, the higher the absolute value of the correlation coefficient and the stronger your linear correlation.

If all points are perfectly on this line, you have a perfect correlation.

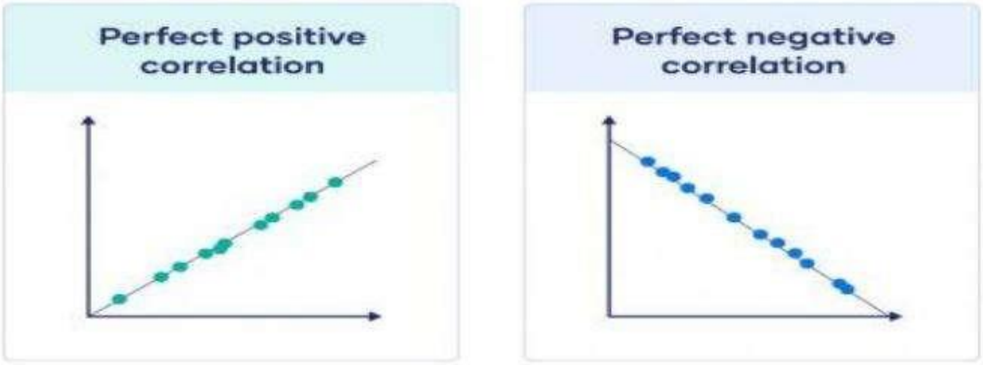


Figure :

If all points are close to this line, the absolute value of your correlation coefficient is high.

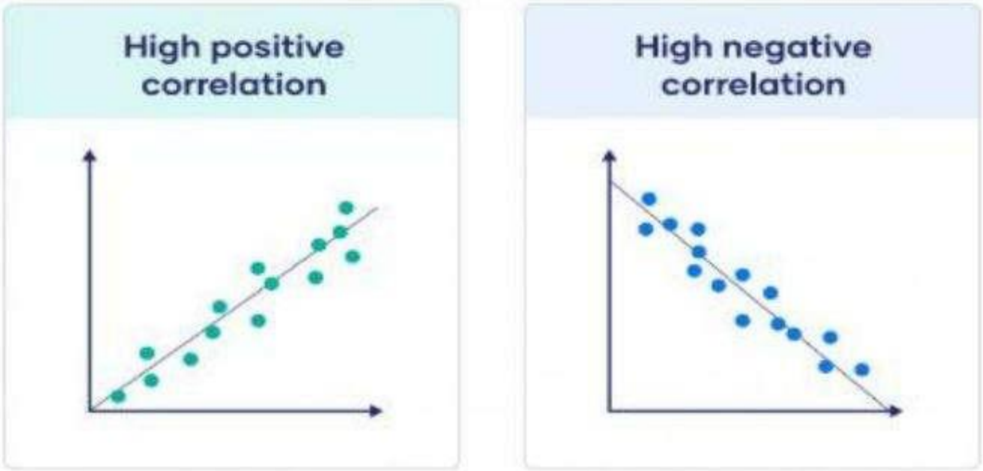


Figure:

If these points are spread far from this line, the absolute value of your correlation coefficient is low.

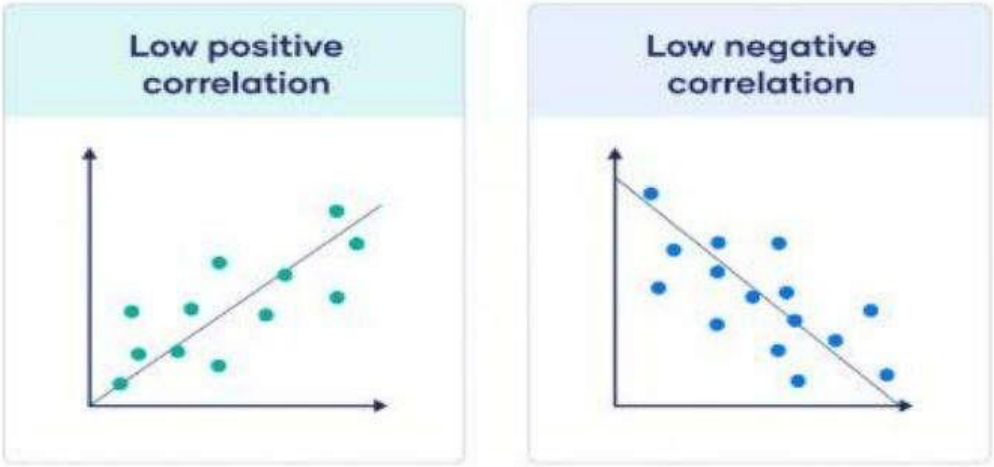
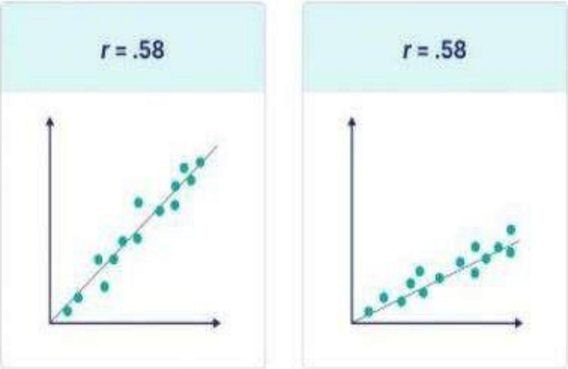


Figure:

Note that the steepness or slope of the line isn't related to the correlation coefficient value. The correlation coefficient doesn't help you predict how much one variable will change based on a given change in the other, because two datasets with the same correlation coefficient value can have lines
with very different slopes.



3.3.1. Types of correlation coefficients

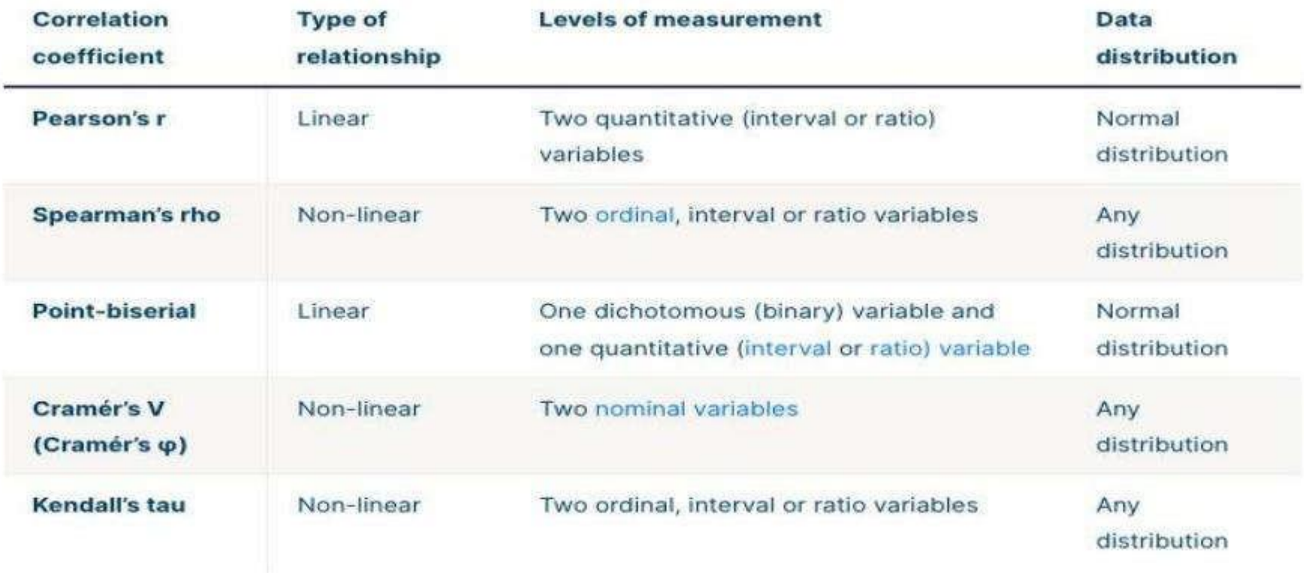
You can choose from many different correlation coefficients based on the linearity of the relationship, the level of measurement of your variables, and the distribution of your data.

For high statistical power and accuracy, it's best to use the correlation coefficient that's most appropriate for your data.

The most commonly used correlation coefficient is Pearson's r because it allows for strong inferences. It's parametric and measures linear relationships. But if your data do not meet all assumptions for this test, you'll need to use a non-parametric test instead.

Non-parametric tests of rank correlation coefficients summarize non-linear relationships between variables. The Spearman's rho and Kendall's tau have the same conditions for use, but Kendall's tau is generally preferred for smaller samples whereas Spearman's rho is more widely used.

The table below is a selection of commonly used correlation coefficients, and we'll cover the two most widely used coefficients in detail in this article.



Correlation coefficient	Type of relationship	Levels of measurement	Data distribution
Pearson's r	Linear	Two quantitative (interval or ratio) variables	Normal distribution
Spearman's rho	Non-linear	Two ordinal, interval or ratio variables	Any distribution
Point-biserial	Linear	One dichotomous (binary) variable and one quantitative (interval or ratio) variable	Normal distribution
Cramér's V (Cramér's φ)	Non-linear	Two nominal variables	Any distribution
Kendall's tau	Non-linear	Two ordinal, interval or ratio variables	Any distribution

Table

3.3.1.1 Pearson's r

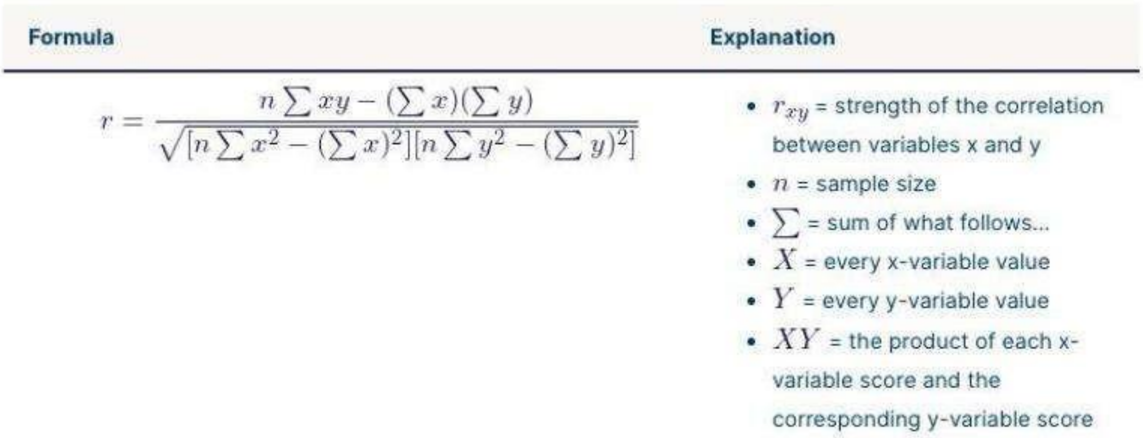
The Pearson's product-moment correlation coefficient, also known as Pearson's r, describes the linear relationship between two quantitative variables.

These are the assumptions your data must meet if you want to use Pearson's r:

- Both variables are on an interval or ratio level of measurement
- Data from both variables follow normal distributions
- Your data have no outliers
- Your data is from a random or representative sample
- You expect a linear relationship between the two variables

The Pearson's r is a parametric test, so it has high power. But it's not a good measure of correlation if your variables have a nonlinear relationship, or if your data have outliers, skewed distributions, or come from categorical variables. If any of these assumptions are violated, you should consider a rank correlation measure.

The formula for the Pearson's r is complicated, but most computer programs can quickly churn out the correlation coefficient from your data. In a simpler form, the formula divides the covariance between the variables by the product of their standard deviations.



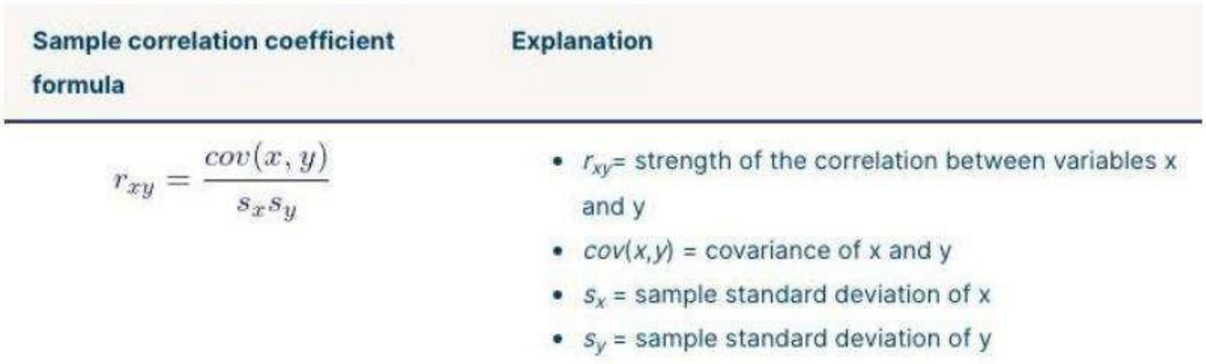
Formula	Explanation
$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$	• r_{xy} = strength of the correlation between variables x and y • n = sample size • $\sum$ = sum of what follows... • X = every x-variable value • Y = every y-variable value • XY = the product of each x-variable score and the corresponding y-variable score

3.3.1.1.1 Pearson sample vs population correlation coefficient formula

When using the Pearson correlation coefficient formula, you'll need to consider whether you're dealing with data from a sample or the whole population.

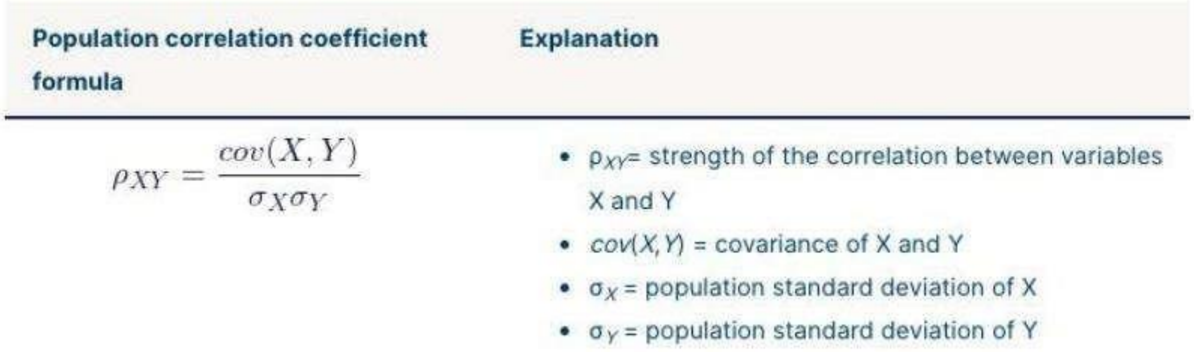
The sample and population formulas differ in their symbols and inputs. A sample correlation coefficient is called r, while a population correlation coefficient is called rho, the Greek letter ρ.

The sample correlation coefficient uses the sample covariance between variables and their sample standard deviations.



Sample correlation coefficient formula	Explanation
$r_{xy} = \frac{cov(x,y)}{s_x s_y}$	• r_{xy}= strength of the correlation between variables x and y • $cov(x,y)$ = covariance of x and y • s_x = sample standard deviation of x • s_y = sample standard deviation of y

The population correlation coefficient uses the population covariance between variables and their population standard deviations.



Population correlation coefficient formula	Explanation
$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$	• ρ_{XY}= strength of the correlation between variables X and Y • $cov(X,Y)$ = covariance of X and Y • σ_X = population standard deviation of X • σ_Y = population standard deviation of Y

3.3.1.2 Spearman's rho

Spearman's rho, or Spearman's rank correlation coefficient, is the most common alternative to Pearson's r. It's a rank correlation coefficient because it uses the rankings of data from each variable (e.g., from lowest to highest) rather than the raw data itself.

You should use Spearman's rho when your data fail to meet the assumptions of Pearson's r. This happens when at least one of your variables is on an ordinal level of measurement or when the data from one or both variables do not follow normal distributions.

While the Pearson correlation coefficient measures the linearity of relationships, the Spearman correlation coefficient measures the monotonicity of relationships.

In a linear relationship, each variable changes in one direction at the same rate throughout the data range. In a monotonic relationship, each variable also always changes in only one direction but not necessarily at the same rate.

Positive monotonic: when one variable increases, the other also increases.
Negative monotonic: when one variable increases, the other decreases.
Monotonic relationships are less restrictive than linear relationships.

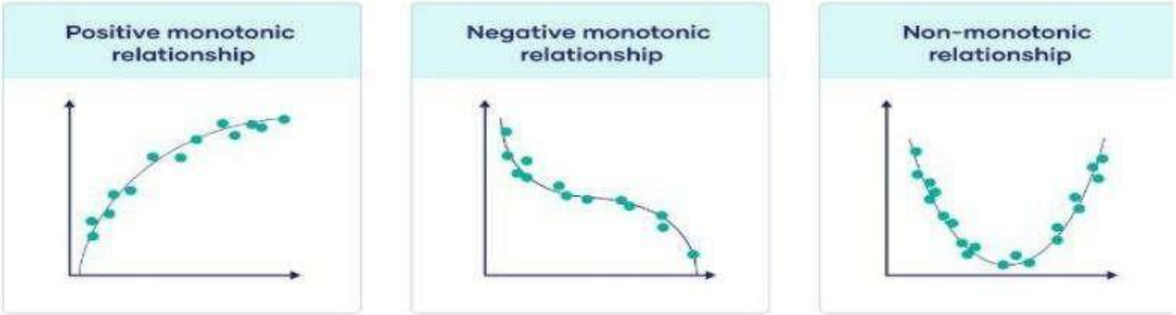


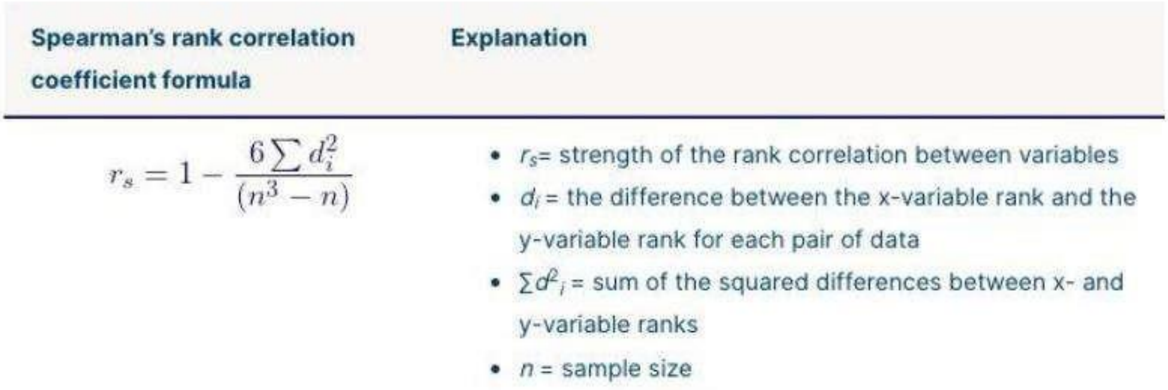
Figure :

3.3.1.2.1 Spearman's rank correlation coefficient formula

The symbols for Spearman's rho are ρ for the population coefficient and rs for the sample coefficient. The formula calculates the Pearson's r correlation coefficient between the rankings of the variable data.

To use this formula, you'll first rank the data from each variable separately from low to high: every datapoint gets a rank from first, second, or third, etc.

Then, you'll find the differences (di) between the ranks of your variables for each data pair and take that as the main input for the formula.



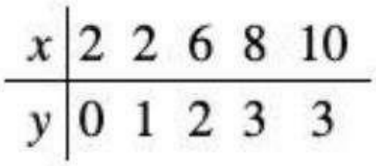
Spearman's rank correlation coefficient formula	Explanation
$r_s = 1 - \frac{6\sum d_i^2}{(n^3 - n)}$	• r_s= strength of the rank correlation between variables • d_i = the difference between the x-variable rank and the y-variable rank for each pair of data • $\sum d^2_i$ = sum of the squared differences between x- and y-variable ranks • n = sample size

If you have a correlation coefficient of 1, all of the rankings for each variable match up for every data pair. If you have a correlation coefficient of -1, the rankings for one variable are the exact opposite of the ranking of the other variable. A correlation coefficient near zero means that there's no monotonic relationship between the variable rankings.

The Least Squares Regression Line

Goodness of Fit of a Straight Line to Data

Once the scatter diagram of the data has been drawn and the model assumptions described in the previous sections at least visually verified (and perhaps the correlation coefficient r computed to quantitatively verify the linear trend), the next step in the analysis is to find the straight line that best fits the data. We will explain how to measure how well a straight line fits a collection of points by examining how well the line y=12x−1 fits the data set



x	2	2	6	8	10
y	0	1	2	3	3

(which will be used as a running example for the next three sections). We will write the equation of this line as yˆ=12x−1 with an accent on the y to indicate that the y-values computed using this equation are not from the data. We will do this with all lines approximating data sets. The line yˆ=12x−1 was selected as one that seems to fit the data reasonably well.

The idea for measuring the goodness of fit of a straight line to data is illustrated in Figure 10.6 "Plot of the Five-Point Data and the Line ", in which the graph of the line yˆ=12x−1 has been superimposed on the scatter plot for the sample data set.

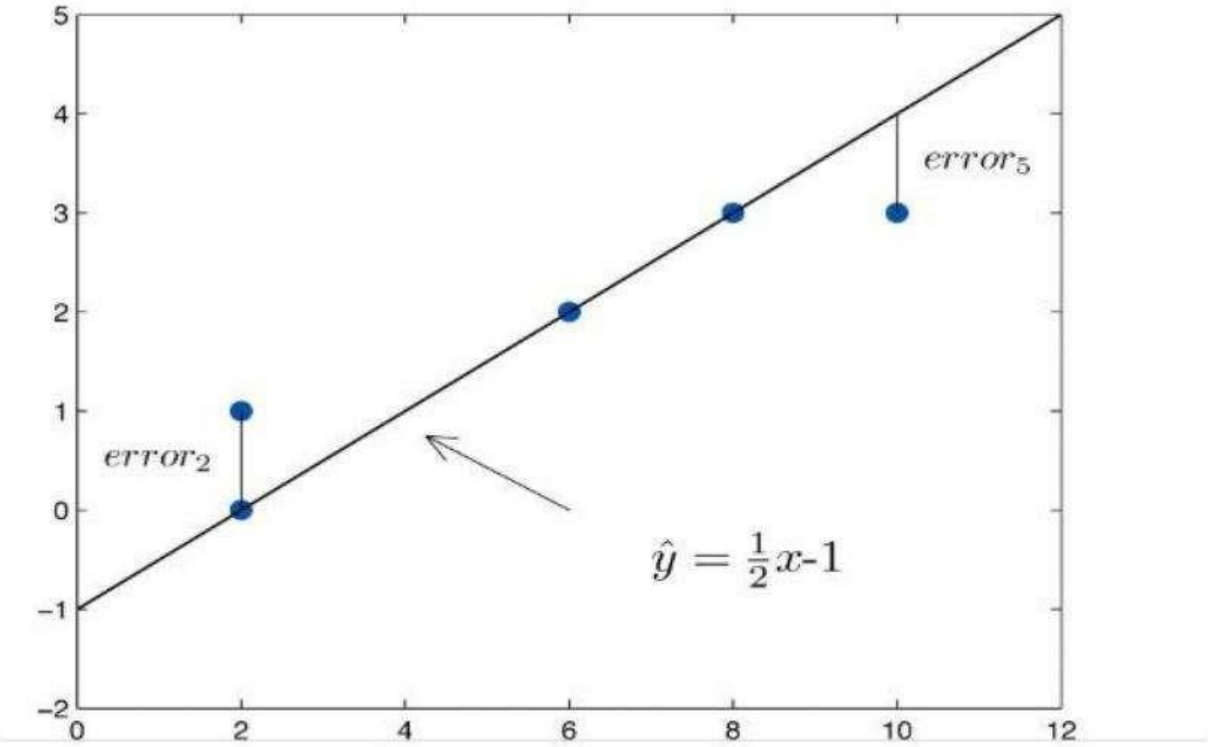
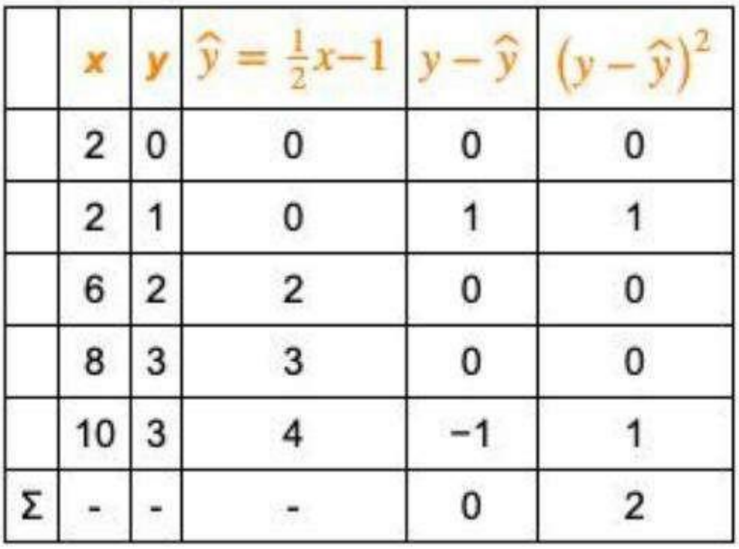


Figure Plot of the Five-Point Data and the Line yˆ=12x−1

To each point in the data set there is associated an “error,” the positive or negative vertical distance from the point to the line: positive if the point is above the line and negative if it is below the line. The error can be computed as the actual y-value of the point minus the y-value yˆ that is “predicted” by inserting the x-value of the data point into the formula for the line:

erroratdatapoint(x,y)=(truey)−(predictedy)=y−yˆ



	x	y	$\hat{y}=\frac{1}{2}x-1$	$y-\hat{y}$	$(y-\hat{y})^2$
	2	0	0	0	0
	2	1	0	1	1
	6	2	2	0	0
	8	3	3	0	0
	10	3	4	−1	1
Σ	-	-	-	0	2

Table The Errors in Fitting Data with a Straight Line

A first thought for a measure of the goodness of fit of the line to the data would be simply to add the errors at every point, but the example shows that this cannot work well in general. The line does not fit the data perfectly (no line can), yet because of cancellation of positive and negative errors the sum of the errors (the fourth column of numbers) is zero. Instead goodness of fit is measured by the sum of the squares of the errors. Squaring eliminates the minus signs, so no cancellation can occur. For the data and line in Figure 10.6 "Plot of the Five-Point Data and the Line " the sum of the squared errors (the last column of numbers) is 2. This number measures the goodness of fit of the line to the data.

Definition

The **goodness of fit** *of a line* yˆ=mx+by^=mx+b *to a set of n pairs* (x,y)(x,y) *of numbers in a sample is the sum of the squared errors*

Σ(y−yˆ)2Σ(y−y^)2

(n terms in the sum, one for each data pair).

The Least Squares Regression Line

Given any collection of pairs of numbers (except when all the x-values are the same) and the corresponding scatter diagram, there always exists exactly one straight line that fits the data better than any other, in the sense of minimizing the sum of the squared errors. It is called the least squares regression line. Moreover there are formulas for its slope and y-intercept.

Definition

Given a collection of pairs (x,y)(x,y) *of numbers (in which not all the x-values are the same), there is a line* yˆ=βˆ1x+βˆ0y^=β^1x+β^0 *that best fits the data in the sense of minimizing the sum of the squared errors. It is called the* least squares regression line. *Its slope* βˆ1β^1 *and y-intercept* βˆ0β^0 *are computed using the formulas*

$$\widehat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \quad \mathit{and} \quad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

Where

$$SS_{xx} = \Sigma x^2 - \frac{1}{n}(\Sigma x)^2, \quad SS_{xy} = \Sigma xy - \frac{1}{n}(\Sigma x)(\Sigma y)$$

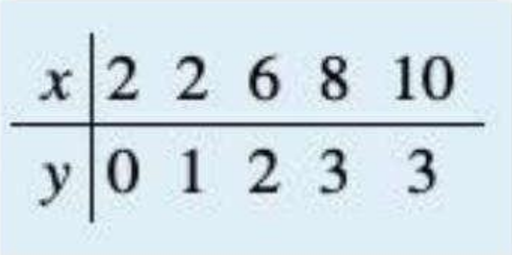
x— —x- *is the mean of all the x-values,* y— —y- *is the mean of all the y-values, and n is the number of pairs in the data set.*

The equation yˆ=βˆ1x+βˆ0y^=β^1x+β^0 *specifying the least squares regression line is called the* least squares regression equation.

Remember from Section 10.3 "Modelling Linear Relationships with Randomness Present" that the line with the equation y=β1x+β0 is called the population regression line. The numbers βˆ1 and βˆ0 are statistics that estimate the population parameters β1 and β

EXAMPLE 1

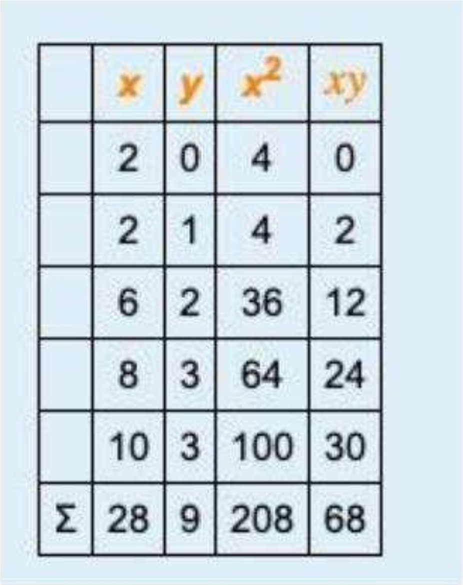
Find the least squares regression line for the five-point data set and verify that it fits the data better than the line y^=12x−1 considered in Section 10.4.1 "Goodness of Fit of a Straight Line to Data".



x	2	2	6	8	10
y	0	1	2	3	3

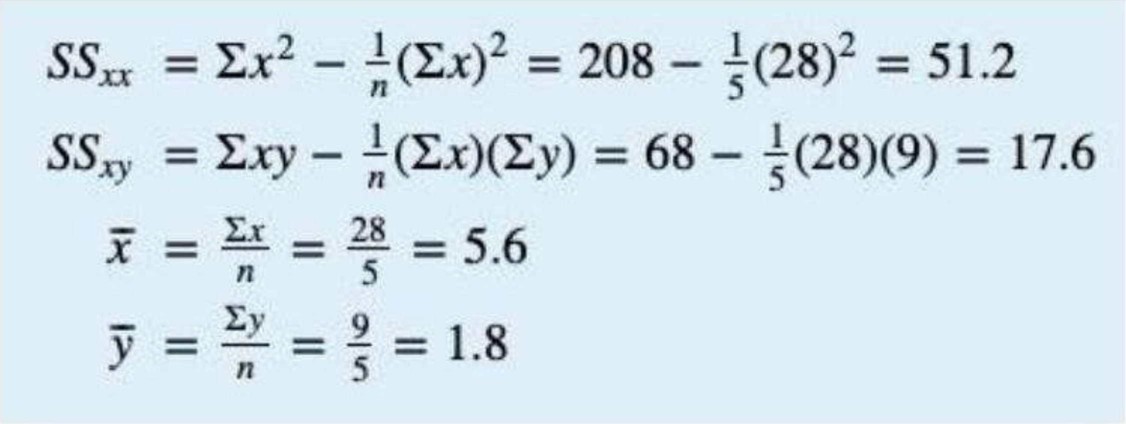
Solution:

In actual practice computation of the regression line is done using a statistical computation package. In order to clarify the meaning of the formulas we display the computations in tabular form.



	x	y	x^2	xy
	2	0	4	0
	2	1	4	2
	6	2	36	12
	8	3	64	24
	10	3	100	30
Σ	28	9	208	68

In the last line of the table we have the sum of the numbers in each column. Using them we compute:



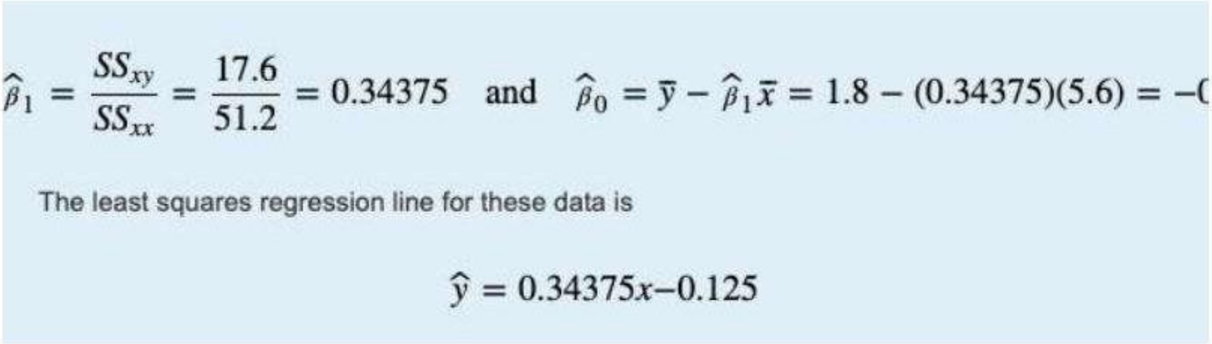
$$SS_{xx} = \Sigma x^2 - \tfrac{1}{n}(\Sigma x)^2 = 208 - \tfrac{1}{5}(28)^2 = 51.2$$

$$SS_{xy} = \Sigma xy - \tfrac{1}{n}(\Sigma x)(\Sigma y) = 68 - \tfrac{1}{5}(28)(9) = 17.6$$

$$\overline{x} = \tfrac{\Sigma x}{n} = \tfrac{28}{5} = 5.6$$

$$\overline{y} = \tfrac{\Sigma y}{n} = \tfrac{9}{5} = 1.8$$

so that



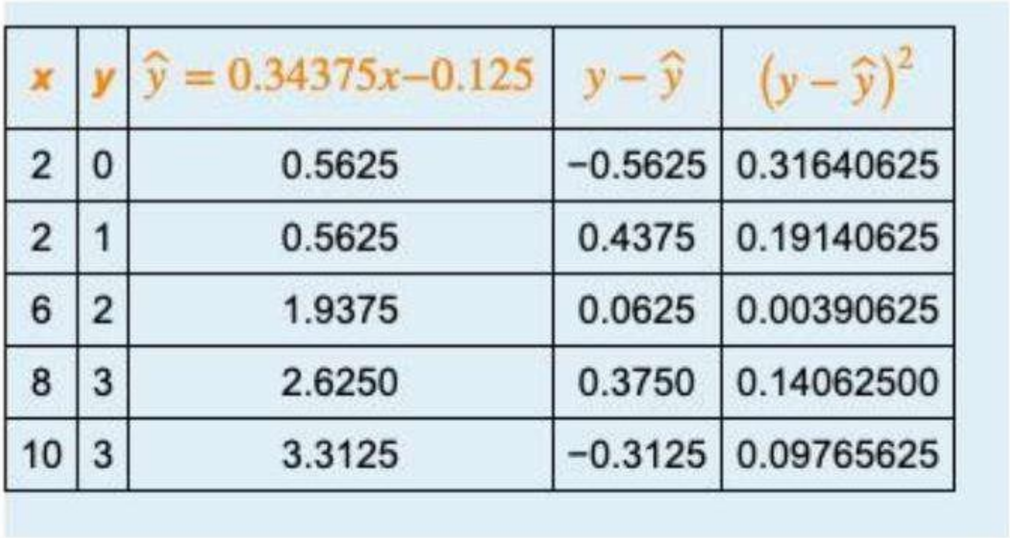
$$\widehat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{17.6}{51.2} = 0.34375 \quad \text{and} \quad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x} = 1.8 - (0.34375)(5.6) = -($$

The least squares regression line for these data is

$$\widehat{y} = 0.34375x - 0.125$$

The computations for measuring how well it fits the sample data are given in Table 10.2 "The Errors in Fitting Data with the Least Squares Regression Line". The sum of the squared errors is the sum of the numbers in the last column, which is 0.75. It is less than 2, the sum of the squared errors for the fit of the line y^=12x−1 to this data set.

THE ERRORS IN FITTING DATA WITH THE LEAST SQUARES REGRESSION LINE



x	y	$\widehat{y} = 0.34375x - 0.125$	$y - \widehat{y}$	$(y - \widehat{y})^2$
2	0	0.5625	−0.5625	0.31640625
2	1	0.5625	0.4375	0.19140625
6	2	1.9375	0.0625	0.00390625
8	3	2.6250	0.3750	0.14062500
10	3	3.3125	−0.3125	0.09765625

3.4 Regression

3.4.1 What is Regression?

- ❖ Define regression with example(2M,8M)
- ❖ Application of regression in real life(2M)

Regression allows researchers to predict or explain the variation in one variable based on another variable.

The variable that researchers are trying to explain or predict is called the response variable. It is also sometimes called the dependent variable because it depends on another variable.

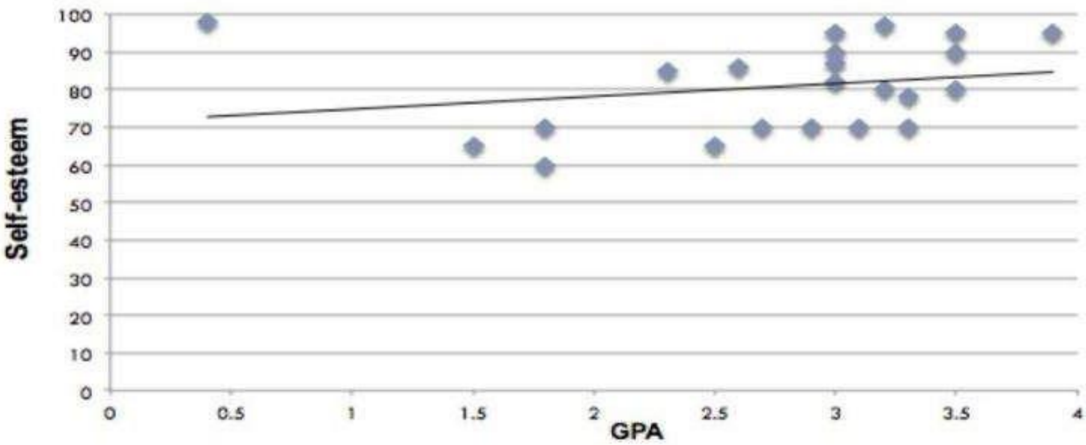
The variable that is used to explain or predict the response variable is called the explanatory variable. It is also sometimes called the independent variable because it is independent of the other variable.

In regression, the order of the variables is very important. The explanatory variable (or the independent variable) always belongs on the x-axis. The response variable (or the dependent variable) always belongs on the y-axis.

Example:

If it is already known that there is a significant correlation between students' GPA and their self-esteem, the next question researchers might ask is: Can students' scores on a self-esteem scale be predicted based on GPA? In other words, does GPA explain self-esteem? These are the types of questions that regression responds to.

**Note that these questions do not imply a causal relationship. In this example, GPA is the explanatory variable (or the independent variable) and self-esteem is the response variable (or the dependent variable). GPA belongs on the x-axis and self-esteem belongs on the y-axis.



Regression is essential for any machine learning problem that involves continuous numbers, which includes a vast array of real-life applications:

1. Financial forecasting, such as estimating housing or stock prices
2. Automobile testing
3. Weather analysis
4. Time series forecasting

3.4.2 Types of Regression

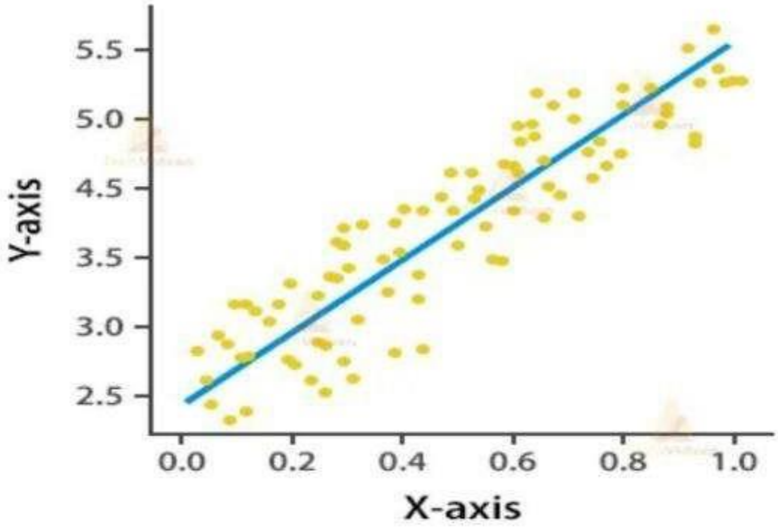
- ❖ Types of regression(2M,16M)
- ❖ What are the three approaches in stepwise regression?(2M)

- ➢ Linear Regression
- ➢ Logistic Regression
- ➢ Polynomial Regression
- ➢ Stepwise Regression
- ➢ Ridge Regression
- ➢ Lasso Regression
- ➢ Elastic Net Regression

3.4.2 .1 LINEAR REGRESSION:

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight.

The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point to the regression line.

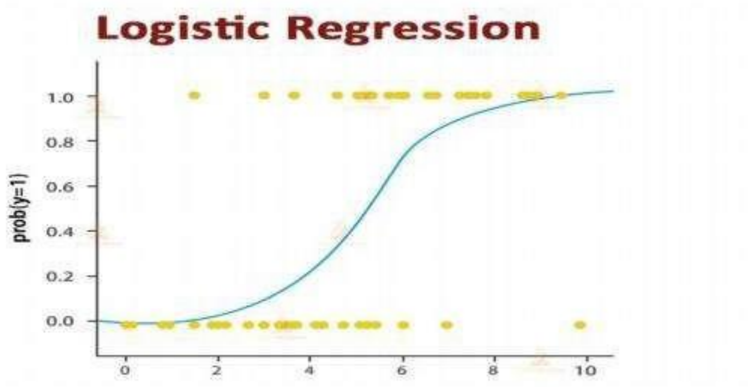


3.4.2.2 LOGISTIC REGRESSION:

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted or not to a particular college. These binary outcomes allow straightforward decisions between two alternatives.

A logistic regression model can take into consideration multiple input criteria. In the case of college acceptance, the logistic function could consider factors such as the student's grade point average, SAT score and number of extracurricular activities. Based on historical data about earlier outcomes involving the same input criteria, it then scores new cases on their probability of falling into one of two outcome categories.

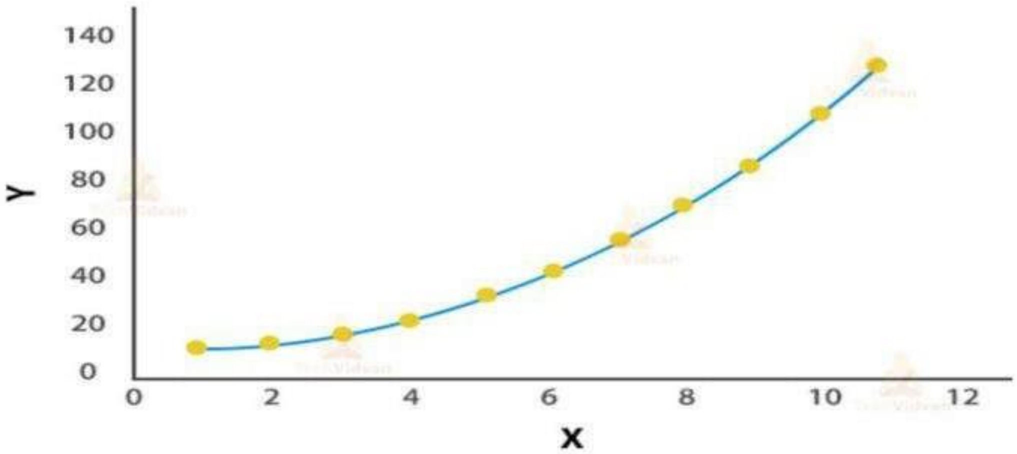


3.4.2.3 POLYNOMIAL REGRESSION:

In a polynomial regression, the power of the independent variable is more than 1. The equation below represents a polynomial equation:

y = a + bx2

In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.



3.4.2.4 STEPWISE REGRESSION:

Stepwise regression is the step-by-step iterative construction of a regression model that involves the selection of independent variables to be used in a final model. It involves adding or removing potential explanatory variables in succession and testing for statistical significance after each iteration.

The availability of statistical software packages makes stepwise regression possible, even in models with hundreds of variables.

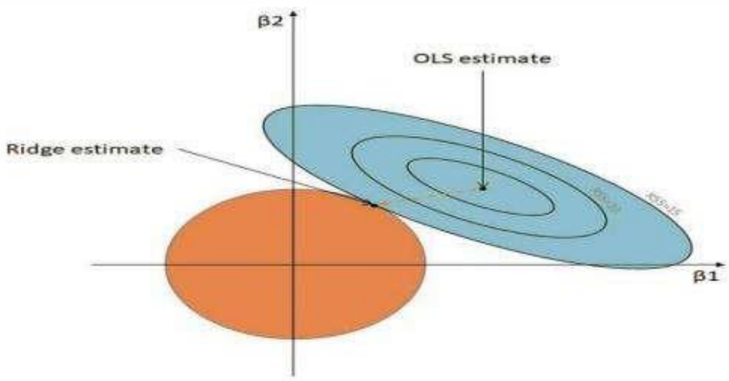
The underlying goal of stepwise regression is, through a series of tests (e.g. F-tests, t-tests) to find a set of independent variables that significantly influence the dependent variable.

there are three approaches to stepwise regression:

- **Forward selection** begins with no variables in the model, tests each variable as it is added to the model, then keeps those that are deemed most statistically significant—repeating the process until the results are optimal.
- **Backward elimination** starts with a set of independent variables, deleting one at a time, then testing to see if the removed variable is statistically significant.
- **Bidirectional elimination** is a combination of the first two methods that test which variables should be included or excluded.

3.4.2.5 RIDGE REGRESSION:

Ridge regression is a type of linear regression technique that is used in machine learning to reduce the overfitting of linear models. Recall that Linear regression is a method of modeling data that represents relationships between a response variable and one or more predictor variables. Ridge regression is used when there are multiple variables that are highly correlated. It helps to prevent overfitting by penalizing the coefficients of the variables. Ridge regression reduces the overfitting by adding a penalty term to the error function that shrinks the size of the coefficients. The penalty term is called the **L2 norm**.Ridge regression is similar to ordinary least squares regression, but the penalty term ensures that the coefficients do not become too large. This can be beneficial when there is a lot of noise in the data, as it prevents the model from being too sensitive to individual data points.



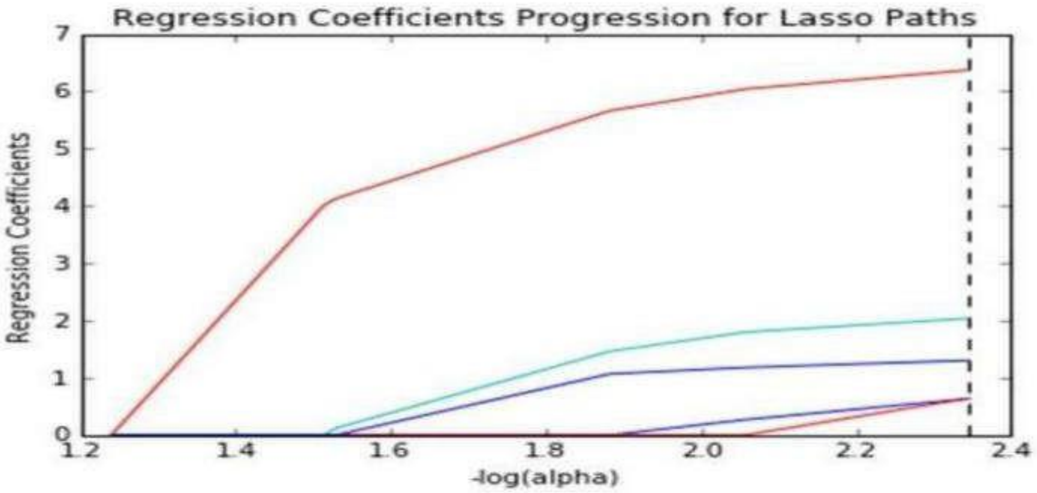
Below is the equation used to denote the Ridge Regression, λ (lambda) resolves the multicollinearity issue:

β = (X^{T}X + λ*I)^{-1}X^{T}y

3.4.2.6 LASSO REGRESSION:

The acronym "LASSO" stands for Least Absolute Shrinkage and Selection Operator.

In short, Lasso Regression is like Ridge Regression regarding its use. However, the only difference is that the data is being fed is not normal. In the case of Lasso Regression, only the required parameters are used, and the rest is made zero. This helps avoid the overfitting in the model. But if independent variables are highly collinear, then Lasso regression chooses only one variable and makes other variables reduce to zero.



3.4.2.7 Elastic Net Regression

Elastic Net regression is being utilized in the case of dominant independent variables being more than one amongst many correlated independent variables.

Also, seasonality & time value factors are made to work together to identify the type of regression.

Elastic Net Regression is a combination of Lasso Regression and Ridge Regression methods. It is prepared with L1 and L2 earlier as regularizer.

The equation represents as:

ElasticNet Regression

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

A clear advantage of trade-off among Lasso and Ridge is that it permits Elastic-Net to acquire a portion of Ridge's dependability under rotation.

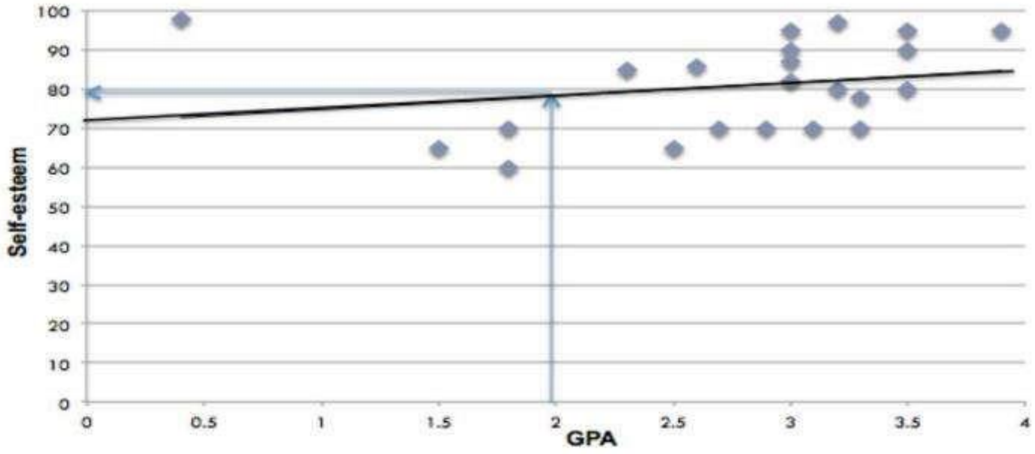
3.4.3 REGRESSION LINE:

- **Define regression lines or why regression lines are important?(2M)**
- **Explain regression line and give an example(13M)**

A regression line is a straight line that describes how a response variable y changes as an explanatory variable x changes. A regression line can be used to predict the value of y for a given value of x.

Regression analysis identifies a regression line. The regression line shows how much and in what direction the response variable changes when the explanatory variable changes. Most individuals in the sample are not located exactly on the line; the line closely approximates all the points. The way this line is computed will be described in more detail later

Example: Predict a student's self-esteem score based on her GPA



- The purpose of a regression line is to make predictions.
- In the above example, it is known how GPA is related to self-esteem. Therefore, if a student's self-esteem has not been measured, but her GPA is known, her self-esteem score can be predicted based on her GPA.
- As an example, if a student has a GPA of 2.0, this score matches up with a score of approximately 78 or 79 on the self-esteem scale. This score has been estimated by looking at the graph.
  - ✓ Draw a straight line up from the point that represents a 2.0 GPA and find where this line intersects with the regression line.
  - ✓ Then, draw a line straight from this point to the self-esteem axis to find the corresponding self-esteem score.

3.4.4 LEAST SQUARE METHOD:

- ❖ **Define least square method(2M)**
- ❖ **What is the formula to calculate Least Square Regression?(6M)**
- ❖ **Define Least Square Regression Line(2M)**
- ❖ **Explain Least Square Regression line With example(13M)**

The least squares method is a form of mathematical regression analysis used to determine the line of best fit for a set of data, providing a visual demonstration of the relationship between the data points. Each point of data represents the relationship between a known independent variable and an unknown dependent variable.

This method of regression analysis begins with a set of data points to be plotted on an x- and y-axis graph. An analyst using the least squares method will generate a line of best fit that explains the potential relationship between independent and dependent variables.

The least squares method is used in a wide variety of fields, including finance and investing. For financial analysts, the method can help to quantify the relationship between two or more variables—such as a stock's share price and its earnings per share (EPS). By performing this type of analysis investors often try to predict the future behavior of stock prices or other factors.

3.4.4.1 FORMULA TO CALCULATE LEAST SQUARE REGRESSION:

The regression line under the Least Squares method is calculated using the following formula –

ŷ = a + bx

Where,

ŷ = dependent variable

x = independent variable

a = y-intercept

b = slope of the line

The slope of line b is calculated using the following formula –

$$b = \frac{\Sigma(x-\overline{x})\,(y-\overline{y})}{\Sigma(x-\overline{x})^2}$$

Or

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Y-intercept, 'a' is calculated using the following formula –

$$a = \frac{\sum y - (b \sum x)}{n}$$

Where,

ŷ = dependent variable

x = independent variable

a = y-intercept

b = slope of the line

The slope of line b is calculated using the following formula –

$$b = \frac{\Sigma(x-\bar{x})\,(y-\overline{y)}}{\Sigma(x-\overline{x)}^{2}}$$

Or

$$b = \frac{\sum xy - \frac{\Sigma x\,\Sigma y}{n}}{\sum x^2 - \frac{(\Sigma x)^2}{n}}$$

Y-intercept, ‘a’ is calculated using the following formula –

$$a = \frac{\sum y - (b \sum x)}{n}$$

3.4.4.2 LEAST SQUARE REGRESSION LINE:

If the data shows a leaner relationship between two variables, the line that best fits this linear relationship is known as a least-squares regression line, which minimizes the vertical distance from the data points to the regression line. The term “least squares” is used because it is the smallest sum of squares of errors, which is also called the "variance."

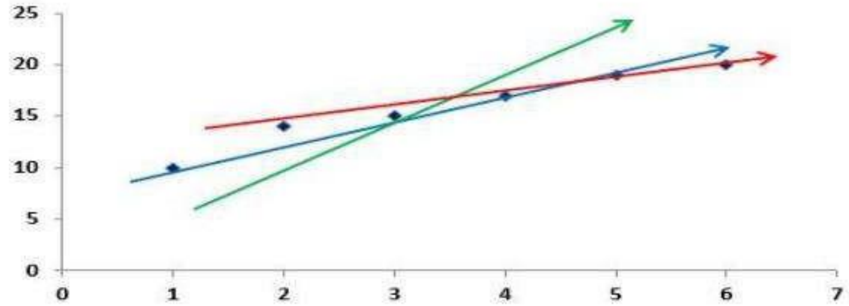
In regression analysis, dependent variables are illustrated on the vertical y-axis, while independent variables are illustrated on the horizontal x-axis. These designations will form the equation for the line of best fit, which is determined from the least squares method.

In contrast to a linear problem, a non-linear least-squares problem has no closed solution and is generally solved by iteration.

EXAMPLE:

The line of best fit is a straight line drawn through a scatter of data points that best represents the relationship between them.

Let us consider the following graph wherein a set of data is plotted along the x and y-axis. These data points are represented using the blue dots. Three lines are drawn through these points – a green, a red, and a blue line. The green line passes through a single point, and the red line passes through three data points. However, the blue line passes through four data points, and the distance between the residual points to the blue line is minimal as compared to the other two lines.



In the above graph, the blue line represents the line of best fit as it lies closest to all the values and the distance between the points outside the line to the line is minimal (i.e., the distance between the residuals to the line of best fit – also referred to as the sums of squares of residuals). In the other two lines, the orange and the green, the distance between the residuals to the lines is greater as compared to the blue line.

3.4.5 STANDARD ERROR OF ESTIMATE:

- ❖ Define Standard Error of Estimate(2M)
- ❖ How Standard Error of Estimate is calculated?(6M)

The **standard error of the estimate** is a way to measure the accuracy of the predictions made by a regression model.

Likewise, a standard deviation which measures the variation in the set of data from its mean, the standard error of estimate also measures the variation in the actual values of Y from the computed values of Y (predicted) on the regression line. It is computed as a standard deviation, and here the deviations are the vertical distance of every dot from the line of average relationship.

Often denoted σest, it is calculated as:

σest = √Σ(y – ŷ)2/n

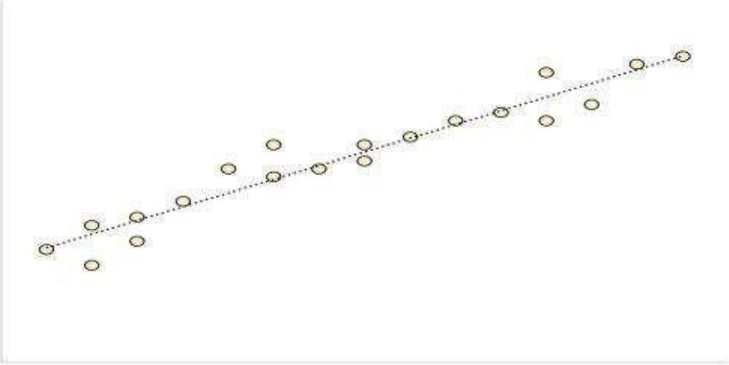
where:

- **y:** The observed value
- **ŷ:** The predicted value
- **n:** The total number of observations

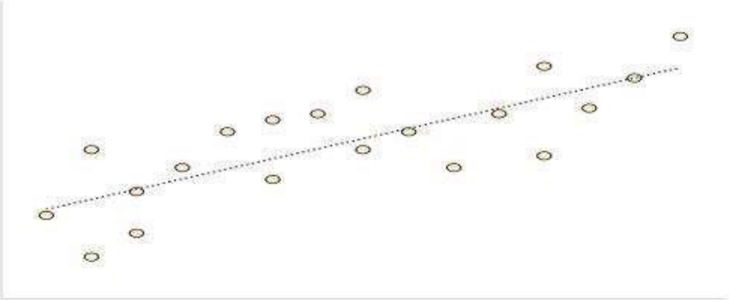
The standard error of the estimate gives us an idea of how well a regression model fits a data set. In particular:

- The smaller the value, the better the fit.
- The larger the value, the worse the fit.

For a regression model that has a small standard error of the estimate, the data points will be closely packed around the estimated regression line:



Conversely, for a regression model that has a large standard error of the estimate, the data points will be more loosely scattered around the regression line:



3.4.6 R-SQUARED:

- Explain R-Squared(2M)
- What is the formula to calculate R-Squared?
- How to interpret R-Squared?(16M)

R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.

After fitting a linear regression model, you need to determine how well the model fits the data.For instance, small R-squared values are not always a problem, and high R-squared values are not necessarily good.

R-squared is always between 0 and 100%:

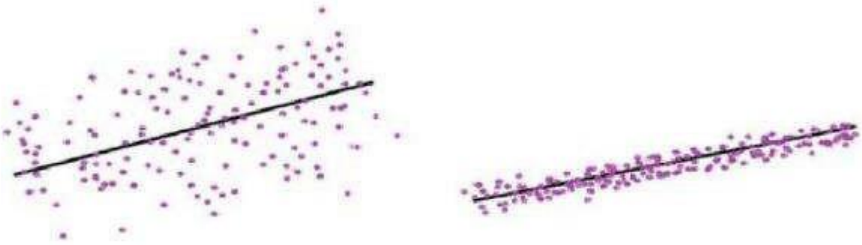
- 0% represents a model that does not explain any of the variation in the response variable around its mean. The mean of the dependent variable predicts the dependent variable as well as the regression model.
- 100% represents a model that explains all the variation in the response variable around its mean.

Usually, the larger the R2, the better the regression model fits your observations.

3.4.6.1 INTERPRETATION OF R2:

Visual Representation of R-squared

You can have a visual demonstration of the plots of fitted values by observed values in a graphical manner. It illustrates how R-squared values represent the scatter around the regression line.



R-squared : 17%

R-squared : 83%

As observed in the pictures above, the value of R-squared for the regression model on the left side is 17%, and for the model on the right is 83%. In a regression model, when the variance accounts to be high, the data points tend to fall closer to the fitted regression line.

However, a regression model with an R2 of 100% is an ideal scenario which is actually not possible. In such a case, the predicted values equal the observed values and it causes all the data points to fall exactly on the regression line.

How to Interpret R squared

The simplest r squared interpretation is how well the regression model fits the observed data values. Let us take an example to understand this.

Consider a model where the R2 value is 70%. Here r squared meaning would be that the model explains 70% of the fitted data in the regression model. Usually, when the R2 value is high, it suggests a better fit for the model.

The correctness of the statistical measure does not only depend on R2 but can depend on other several factors like the nature of the variables, the units on which the variables are measured, etc. So, a high R-squared value is not always likely for the regression model and can indicate problems too.

A low R-squared value is a negative indicator for a model in general. However, if we consider the other factors, a low R2 value can also end up in a good predictive model.

Calculation of R-squared

R- squared can be evaluated using the following formula:

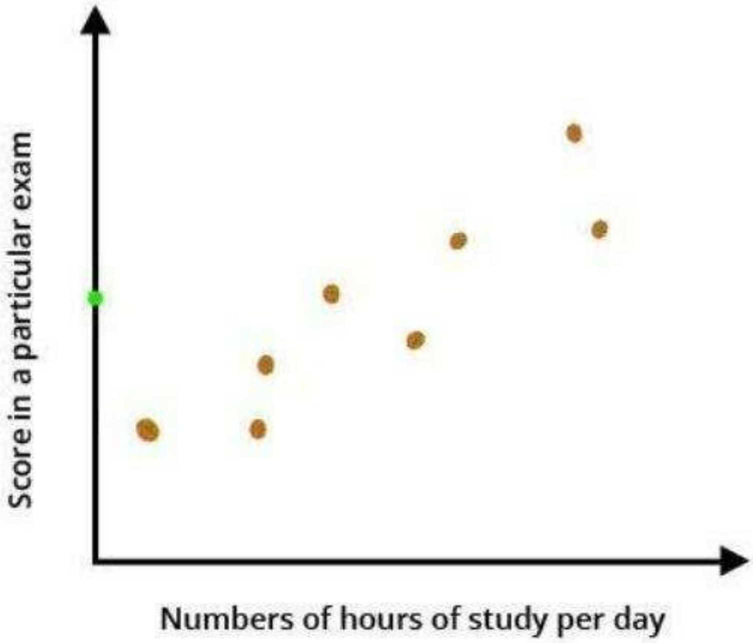
$$\text{R-squared} = \frac{SS_{\text{regresion}}}{SS_{\text{total}}}$$

Where:

- SSregression – Explained sum of squares due to the regression model.
- SStotal – The total sum of squares.

The sum of squares due to regression assesses how well the model represents the fitted data and the total sum of squares measures the variability in the data used in the regression model.

Now let us come back to the earlier situation where we have two factors: number of hours of study per day and the score in a particular exam to understand the calculation of R-squared more effectively. Here, the target variable is represented by the score and the independent variable by the number of hours of study per day.

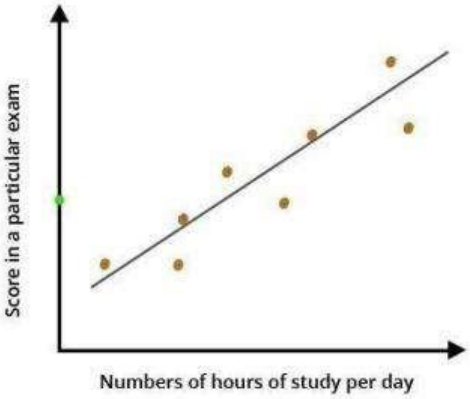


In this case, we will need a simple linear regression model and the equation of the model will be as follows:

$$\hat{y} = w1x1 + b$$

The parameters w1 and b can be calculated by reducing the squared error over all the data points. The following equation is called the least square function:

$$\text{minimize} \sum(yi - w1x1i - b)2$$



Now, to calculate the goodness-of-fit, we need to calculate the variance:

$$var(u) = 1/n\sum(ui - \bar{u})2$$

where, n represents the number of data points.

Now, R-squared calculates the amount of variance of the target variable explained by the model, i.e. function of the independent variable.

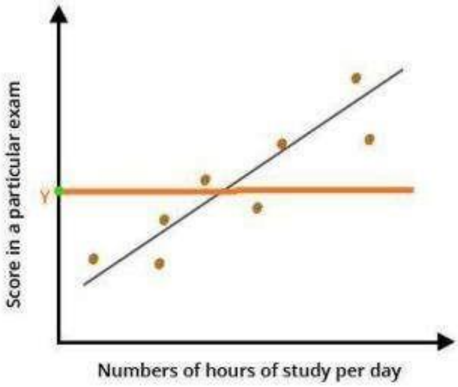
However, in order to achieve that, we need to calculate two things:

- Variance of the target variable:

$var(avg) = \sum(yi - \bar{Y})2$

- Variance of the target variable around the best-fit line:

$var(model) = \sum(yi - \hat{y})2$



Finally, we can calculate the equation of R-squared as follows:

$R2 = 1 - [var(model)/var(avg)] = 1 -[\sum(yi - \hat{y})2/\sum(yi - \bar{Y})2]$

3.4.7 MULTIPLE REGRESSION:

- ❖ Explain Multiple regression?(2M)
- ❖ Explain linear regression and multiple regression equation with example(13M)
- ❖ Assumptions of Multiple Regression Equations(2M)
- ❖ Benefits of Multiple Regression Equations(2M)

Multiple regression is a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value. Each predictor value is weighed, the weights denoting their relative contribution to the overall prediction.

$Y=a+b1X1+b2X3+...+bnXn$

Here Y is the dependent variable, and X1,…,Xn are the *n* independent variables. In calculating the weights, a, b1,…,bn, regression analysis ensures maximal prediction of the dependent variable from the set of independent variables. This is usually done by least squares estimation.

In the case of linear regression, although it is used commonly, it is limited to just one independent and one dependent variable. Apart from that, linear regression restricts the training data set and does not predict a non-linear regression.

For the same limitations and to cover them, we use multiple regression. It focuses on overcoming one particular limitation and that is allowing to analyze more than one independent variable.

3.4.7.1 Multiple regression equation

We will start the discussion by first taking a look at the linear regression equation:

y = bx + a

Where,

y is a dependent variable we need to find, x is an independent variable. The constants a and b drive the equation. But according to our definition, as the multiple regression takes several independent variables (x), so for the equation we will have multiple x values too:

y = b1x1 + b2x2 + … bnxn + a

Here, to calculate the value of the dependent variable y, we have multiple independent variables x1, x2, and so on. The number of independent variables can grow till n and the constant b with every variable denotes its numeric value. The purpose of the constant a is to denote the dependent variable's value in case when all the independent variable values turn to zero.

Example: A researcher decides to study students' performance at a school over a period of time. He observed that as the lectures proceed to operate online, the performance of students started to decline as well. The parameters for the dependent variable "decrease in performance" are various independent variables like "lack of attention, more internet addiction, neglecting studies" and much more.

So for the above example, the multiple regression equation would be:

y = b1 * attention + b2 * internet addiction + b3 * technology support + … bnxn + a

3.4.7.2 ASSUMPTIONS OF MULTIPLE REGRESSION ANALYSIS:

- The variables considered for the model should be relevant and the model should be reliable.
- The model should be linear and not non-linear.
- Variables must have a normal distribution
- The variance should be constant for all levels of the predicted variable.

3.4.7.3 BENEFITS OF MULTIPLE REGRESSION ANALYSIS:

- Multiple regression analysis helps us to better study the various predictor variables at hand.

- ✧ It increases reliability by avoiding dependency on just one variable and having more than one independent variable to support the event.
- ✧ Multiple regression analysis permits you to study more formulated hypotheses that are possible.

3.4.8 REGRESSION TOWARDS THE MEAN:

- ❖ Define regression towards mean(2M)
- ❖ Explain regression towards mean with example(13M)

In statistics, **regression toward the mean** (also called **reversion to the mean**, and **reversion to mediocrity**) is a concept that refers to the fact that if one sample of a random variable is extreme, the next sampling of the same random variable is likely to be closer to its mean. Furthermore, when many random variables are sampled and the most extreme results are intentionally picked out, it refers to the fact that (in many cases) a second sampling of these picked-out variables will result in "less extreme" results, closer to the initial mean of all of the variables.

Regression to the mean usually happens because of sampling error. A good sampling technique is to randomly sample from the population. If you don't (i.e. if you asymmetrically sample), then your results may be abnormally high or low for the average and therefore would regress back to the mean. Regression to the mean can also happen because you take a very small, unrepresentative sample (say, the highest 1 percent of the population or the lowest ten percent).

Formula for the Percent of Regression to the Mean:

You can use the following formula to find the percent for any set of data:

Percent of Regression to the Mean = 100(1-r)

where r is the correlation coefficient.

Why 1-r?
Note: In order to understand this discussion you should be very familiar with r, the correlation coefficient.

The percent of regression to the mean takes into account the correlation between the variables. Take two extremes:
If r=1 (i.e. perfect correlation), then 1-1 = 0 and the regression to the mean is zero. In other words, if your data has perfect correlation, it will never regress to the mean.
With an r of zero, there is 100 percent regression to the mean. In other words, data with an r of zero will *always* regress to the mean.

EXAMPLE:

If your favorite team won the championship last year, what does that mean for their chances for winning next season? This is an important question, often with money or pride on the line (The League, anyone?). To the extent this is due to skill (the team is in good condition, top coach etc.), their win signals that it's more likely they'll win next year. But the greater the extent this is due to luck (other teams embroiled in a drug scandal, favourable draw, draft picks turned out well etc.), the less likely it is they'll win next year. This is because of the statistical concept of regression to the mean.

Another example,

because of regression toward the mean, we would expect that students who made the top five scores on the first statistics exam would not make the top five scores on the second statistics exam. Although all five students might score above the mean on the second exam, some of their scores would regress back toward the mean. Most likely, the top five scores on the first exam reflect two components. One relatively permanent component reflects the fact that these students are superior because of good study habits, a strong aptitude for quantitative reasoning, and so forth. The other relatively transitory component reflects the fact that, on the day of the exam, at least some of these students were very lucky because all sorts of little chance factors, such as restful sleep, a pleas ant commute to campus, etc., worked in their favor. On the second test, even though the scores of these five students continue to reflect an above-average permanent component, some of their scores will suffer because of less good luck or even bad luck. The net effect is that the scores of at least some of the original five top students will drop below the top five scores—that is, regress *back* toward the mean—on the second exam. (When signifi cant regression toward the mean occurs after a spectacular performance by, for example, a rookie athlete or a first-time author, the term *sophomore jinx* often is invoked.)

There is good news for those students who made the five lowest scores on the first exam. Although all five students might score below the mean on the second exam, some of their scores probably will regress *up* toward the mean. On the second exam, some of them will not be as unlucky. The net effect is that the scores of at least some of the original five lowest scoring students will move above the bottom five scores—that is, regress up toward the mean—on the second

UNIT III INFERENTIAL STATISTICS

Populations – samples – random sampling – Sampling distribution- standard error of the mean - Hypothesis testing – z-test – z-test procedure –decision rule – calculations – decisions – interpretations - one-tailed and two-tailed tests – Estimation – point estimate – confidence interval – level of confidence – effect of sample size.

3.1 POPULATIONS

Any complete set of observations (or potential observations) may be characterized as a ***Population.*** Accurate descriptions of populations specify the nature of the observations to be taken. For example, a population might be described as "attitudes toward abortion of currently enrolled students at Bucknell University" or as "SAT critical reading scores of currently enrolled students at Rutgers University".

1. **Real Populations**

Pollsters, such as the Gallup Organization, deal with real populations. A *real* population is one in which all potential observations are accessible at the time of sampling. Examples of real populations, the ages of all visitors to Disneyland on a given day, the ethnic backgrounds of all current employees of the U.S. Postal Department, and presidential preferences of all currently registered voters in the United States. Incidentally, federal law requires that a complete survey be taken every 10 years of the real population of all U.S. house- holds at considerable expense, involving thousands of data collectors as a means of revising election districts for the House of Representatives. (An estimated undercount of millions of people, particularly minorities, in both the 2000 and 2010 censuses has revived a suggestion, long endorsed by statisticians, that the entire U.S. population could be estimated more accurately if a highly trained group of data collectors focused only on a random sample of households.).

2. **Hypothetical Populations**

A *hypothetical* population is one in which all potential observations are not accessible at the time of sampling. In most experiments, subjects are selected from very small, uninspiring real populations: the lab rats housed in the local animal colony or student volunteers from general psychology classes. Experimental subjects often are viewed, nevertheless, as a sample from a much larger hypothetical population, loosely described as "the scores of all similar animal subjects (or student volunteers) who could conceivably undergo the present experiment." According to the rules of inferential statistics, generalizations should be made only to real populations that, in fact, have been sampled. Generalizations to hypothetical populations should be viewed, therefore, as provisional conclusions based on the wisdom of the researcher rather than on any logical or statistical necessity. In effect, it's an open question often answered only by additional experimentation whether or not a given experimental finding merits the generality assigned to it by the researcher.

3.1.2 SAMPLES

Any subset of observations from a population may be characterized as a **sample**. In typical applications of inferential statistics, the sample size is small relative to the population size. For example, less than 1 percent of all U.S. worksites are included in the Bureau of Labor Statistics' monthly survey to estimate the rate of unemployment. And although, only 1475 likely voters had been sampled in the final poll for the 2012 presidential election by the NBC News/*Wall Street Journal*, it correctly predicted that Obama would be the slim winner of the popular vote.

Optimal Sample Size

There is no simple rule of thumb for determining the best or optimal sample size for any particular situation. Often sample sizes are in the hundreds or even the thousands for surveys, but they are less than 100 for most experiments. Optimal sample size depends on the answers to a number of questions, including "What is the estimated variability among observations?" and "What is an acceptable amount of error in our conclusion?" Once these types of questions have been answered that is the result, the specific procedures can be followed to determine the optimal sample size for any situation.

3.1.3 RANDOM SAMPLING

The valid use of techniques from inferential statistics requires that samples be random.

***Random sampling* occurs if, at each stage of sampling, the selection process guarantees that all potential observations in the population have an equal chance of being included in the sample.**

It's important to note that randomness describes the *selection process* that is, the conditions under which the sample is taken and not the particular pattern of observations in the sample. Having established that sampling is random, you still can't predict anything about the unique pattern of observations in that sample. The observations in the sample should be representative of those in the population, but there is no guarantee that they actually will be.

Casual or Haphazard, Not Random

A casual or haphazard sample doesn't qualify as a random sample. Not every student at UC San Diego has an equal chance of being sampled if, for instance, a pollster casually selects only students who enter the student union. Obviously excluded from this sample are all those students (few, we hope) who never enter the student union. Even the final selection of students from among those who do enter the student union might reflect the pollster's various biases, such as an unconscious preference for attractive students who are walking alone.

3.2 Sampling distribution

WHAT IS A SAMPLING DISTRIBUTION?

Random samples rarely represent the underlying population exactly. Even a mean math score of 533 could originate, just by chance, from a population of freshmen whose mean equals the national average of 500. Accordingly, generalizations from a single sample to a population are much more tentative. Indeed, generalizations are based not merely on the single sample mean of 533 but also on its distribution a distribution of sample means for all possible random samples. Representing the statistician's model of random outcomes,

The *sampling distribution of the mean* refers to the probability distribution of means for all possible random samples of a given size from some population.

In effect, this distribution describes the variability among sample means that could occur just by chance and thereby serves as a frame of reference for generalizing from a single sample mean to a population mean.

The sampling distribution of the mean allows us to determine whether, given the variability among all possible sample means, the one observed sample mean can be viewed as a *common* outcome or as a *rare* outcome (from a distribution centered, in this case, about a value of 500). If the sample mean of 533 qualifies as a *common* outcome in this sampling distribution, then the difference between 533 and 500 isn't large enough, relative to the variability of all possible sample means, to signify that anything special is happening in the underlying population. Therefore, we can conclude that the mean math score for the entire freshman class could be the same as the national average of 500. On the other hand, if the sample mean of 533 qualifies as a *rare* outcome in this sampling distribution, then the difference between 533 and 500 is large enough, relative to the variability of all possible sample means, to signify that something special probably is happening in the underlying population. Therefore, we can conclude that the mean math score for the entire freshman class probably exceeds the national average of 500.

All Possible Random Samples

When attempting to generalize from a single sample mean to a population mean, must consult the sampling distribution of the mean. In the present case, this distribution is based on *all possible* random samples, each of size 100 that can be taken from the local population of freshmen. *All possible random samples* refers not to the number of samples of size 100 required to *survey completely* the local population of freshmen but to the number of different ways in which a *single* sample of size 100 can be selected from this population.

"All possible random samples" tends to be a huge number. For instance, if the local population contained at least 1,000 freshmen, the total number of possible random samples, each of size 100, would be astronomical in size. The 301 digits in this number would dwarf even the national debt. Even with the aid of a computer, it would be a horrendous task to construct this sampling distribution from scratch, itemizing each mean for all possible random samples.

Fortunately, statistical theory supplies us with considerable information about the sampling distribution of the mean, as will be discussed in the remainder of this chapter. Armed with this information about sampling distributions, we'll return to the current example in the next chapter and test the claim that the mean math score for the local population of freshmen equals the national average of 500. Only at that point and not at the end of this chapter should you expect to understand completely the role of sampling distributions in practical applications.

3.2.1 CREATING A SAMPLING DISTRIBUTION FROM SCRATCH

Let’s establish precisely what constitutes a sampling distribution by creating one from scratch under highly simplified conditions. Imagine some ridiculously small population of four observations with values of 2, 3, 4, and 5, as shown in Figure 9.1**.** Next, itemize all possible random samples, each of size two, that could be taken from this population. There are four possibilities on the first draw from the population and also four possibilities on the second draw from the population, as indicated in Table 9.1**.*** The two sets of possibilities combine to yield a total of 16 possible samples. At this point, remember, we’re clarifying the notion of a sampling distribution of the mean. In practice, only a single random sample, not 16 possible samples, would be taken from the population; the sample size would be very small relative to a much larger population size, and, of course, not all observations in the population would be known.

For each of the 16 possible samples, Table 9.1 also lists a sample mean (found by adding the two observations and dividing by 2) and its probability of occurrence (expressed as ^{1}/16, since each of the 16 possible samples is equally likely). When cast into a relative frequency or probability distribution, as in Table 9.2**,** the 16 sample means constitute the sampling distribution of the mean, previously defined as the probability distribution of means for all possible random samples of a given size from some population. Not all values of the sample mean occur with equal probabilities in Table 9.2 since some values occur more than once among the 16 possible samples. For instance, a sample mean value of 3.5 appears among 4 of 16 possibilities and has a probability of ^{4}/16.

1. Probability of a Particular Sample Mean

The distribution in Table 9.2 can be consulted to determine the probability of obtaining a particular sample mean or set of sample means. For example, the probability of a randomly selected sample mean of 5.0 equals ^{1}/16 or .0625. According to the addition rule for mutually exclusive outcomes, the probability of a ran domly selected sample mean of either 5.0 or 2.0 equals 1 /16 + 1 /16 = 2 /16 = .1250.

Table 9.1
ALL POSSIBLE SAMPLES OF SIZE TWO FROM A MINIATURE POPULATION

	ALL POSSIBLE SAMPLES	MEAN ($\overline{X}$)	PROBABILITY
(1)	2,2	2.0	$\frac{1}{16}$
(2)	2,3	2.5	$\frac{1}{16}$
(3)	2,4	3.0	$\frac{1}{16}$
(4)	2,5	3.5	$\frac{1}{16}$
(5)	3,2	2.5	$\frac{1}{16}$
(6)	3,3	3.0	$\frac{1}{16}$
(7)	3,4	3.5	$\frac{1}{16}$
(8)	3,5	4.0	$\frac{1}{16}$
(9)	4,2	3.0	$\frac{1}{16}$
(10)	4,3	3.5	$\frac{1}{16}$
(11)	4,4	4.0	$\frac{1}{16}$
(12)	4,5	4.5	$\frac{1}{16}$
(13)	5,2	3.5	$\frac{1}{16}$
(14)	5,3	4.0	$\frac{1}{16}$
(15)	5,4	4.5	$\frac{1}{16}$
(16)	5,5	5.0	$\frac{1}{16}$

Table 9.2
SAMPLING DISTRIBUTION OF THE MEAN (SAMPLES OF SIZE TWO FROM A MINIATURE POPULATION)

SAMPLE MEAN ($\overline{X}$)	PROBABILITY
5.0	$\frac{1}{16}$
4.5	$\frac{2}{16}$
4.0	$\frac{3}{16}$
3.5	$\frac{4}{16}$
3.0	$\frac{3}{16}$
2.5	$\frac{2}{16}$
2.0	$\frac{1}{16}$

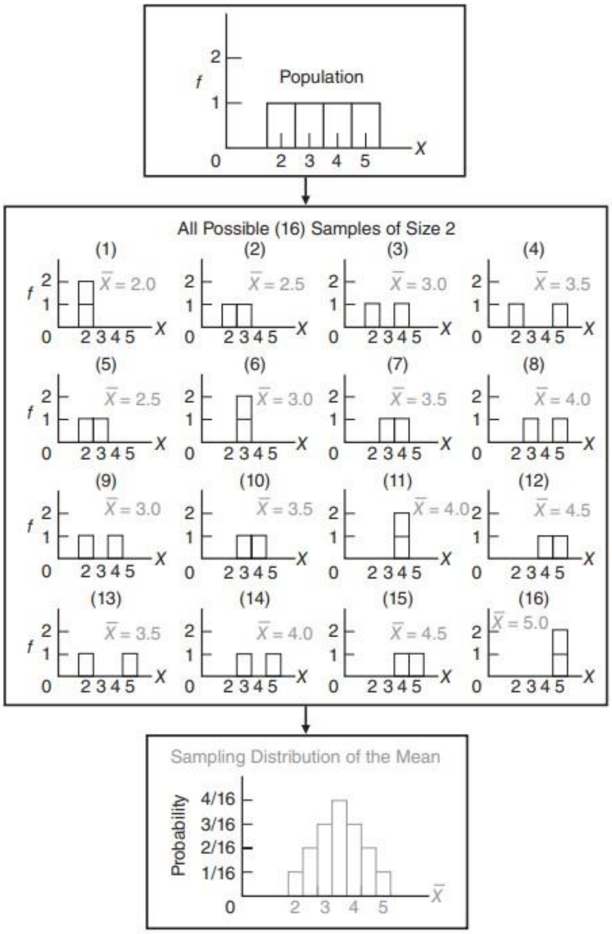


FIGURE 9.2
Emergence of the sampling distribution of the mean from all possible samples.

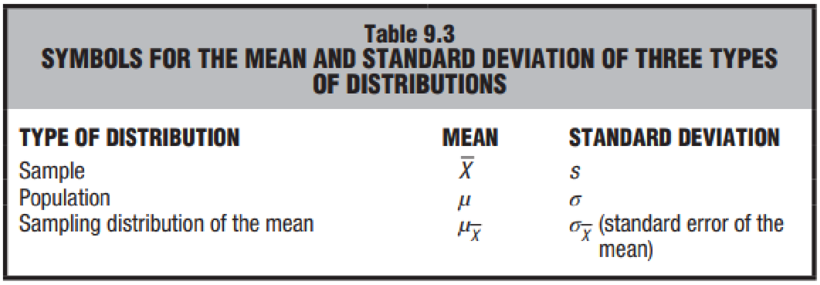


Table 9.3

SYMBOLS FOR THE MEAN AND STANDARD DEVIATION OF THREE TYPES OF DISTRIBUTIONS

TYPE OF DISTRIBUTION	MEAN	STANDARD DEVIATION
Sample	$\bar{X}$	s
Population	μ	σ
Sampling distribution of the mean	$\mu_{\bar{X}}$	$\sigma_{\bar{X}}$ (standard error of the mean)

3.2.3 MEAN OF ALL SAMPLE MEANS ($\mu_{\bar{X}}$)

The distribution of sample means itself has a mean. The mean of the sampling distribution of the mean always equals the mean of the population.

Expressed in symbols,

MEAN OF THE SAMPLING DISTRIBUTION

$$\mu_{\bar{X}} = \mu \qquad (9.1)$$

Where $\mu_{\bar{X}}$ represents the mean of the sampling distribution and μ represents the mean of the population.

1. **Interchangeable Means**

The mean of all sample means ($\mu_{\bar{X}}$) always equals the mean of the population (μ), these two terms are interchangeable in inferential statistics. Any claims about thepopulation mean can be transferred directly to the mean of the sampling distribution, and vice versa. If, as claimed, the mean math score for the local population of freshmen equals the national average of 500, then the mean of the sampling distribution also automatically will equal 500. For the same reason, it's permissible to view the one observed sample mean of 533 as a deviation either from the mean of the sampling distribution or from the mean of the population. It should be apparent, therefore, that whether an expression involves $\mu_{\bar{X}}$ or μ, it reflects, at most, a difference in emphasis on either the sampling distribution or the population, respectively, rather than any difference in numerical value.

Explanation

Although important, it's not particularly startling that the mean of all sample means equals the population mean. As can be seen in Figure 9.2, samples are not exact replicas of the population, and most sample means are either larger or smaller than the population mean (equal to 3.5 in Figure 9.2). By taking the mean of all sample means, however, you effectively neutralize chance differences between sample means and retain a value equal to the population mean.

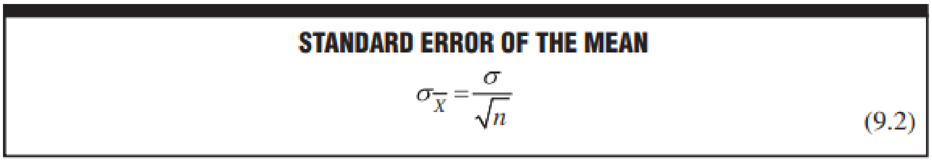
3.2.4 STANDARD ERROR OF THE MEAN ($\sigma_{\bar{X}}$)

The distribution of sample means also has a standard deviation, referred to as the standard error of the mean.

The standard error of the mean equals the standard deviation of the population divided by the square root of the sample size.

1. **STANDARD ERROR OF THE MEAN ($\sigma_{\bar{X}}$)**

Expressed in symbols,



STANDARD ERROR OF THE MEAN

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \quad (9.2)$$

where $\sigma_{\overline{X}}$ represents the standard error of the mean; σ represents the standard deviation of the population; and n represents the sample size.

2. **Special Type of Standard Deviation**

The standard error of the mean serves as a special type of standard deviation that measures variability in the sampling distribution. It supplies us with a standard, much like a yardstick, that describes the amount by which sample means deviate from the mean of the sampling distribution or from the population mean. The error in standard error refers not to computational errors, but to errors in generalizations attributable to the fact that, just by chance, most random samples aren't exact replicas of the population.

The standard error of the mean as a rough measure of the average amount by which sample means deviate from the mean of the sampling distribution or from the population mean.

Insofar as the shape of the distribution sample means approximates a normal curve, as described in the next section, about 68 percent of all sample means deviate less than one standard error from the mean of the sampling distribution, whereas only about 5 percent of all sample means deviate more than two standard errors from the mean of this distribution.

3. **Effect of Sample Size**

A most important implication of Formula 9.2 is that whenever the sample size equals two or more, the variability of the sampling distribution is less than that in the population. A modest demonstration of this effect appears in Figure 9.2, where the means of all possible samples cluster closer to the population mean (equal to 3.5) than do the four original observations in the population. A more dramatic demonstration occurs with larger sample sizes. Earlier in this chapter, for instance, 110 was given as the value of σ, the population standard deviation for SAT scores. Much smaller is the variability in the sampling distribution of mean SAT scores, each based on samples of 100 freshmen. According to Formula 9.2, in the present example,

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{110}{\sqrt{100}} = \frac{110}{10} = 11$$

there is a tenfold reduction in variability, from 110 to 11, when our focus shifts from the population to the sampling distribution.

According to Formula 9.2, any increase in sample size translates into a smaller standard error and, therefore, into a new sampling distribution with less variability. With a larger sample size, sample means cluster more closely about the mean of the sampling distribution and about the mean of the population and, therefore, allow more precise generalizations from samples to populations.

3.2.5 SHAPE OF THE SAMPLING DISTRIBUTION

A product of statistical theory, expressed in its simplest form, **the central limit theorem states that, regardless of the shape of the population, the shape of the sampling distribution of the mean approximates a normal curve if the sample size is sufficiently large.**

According to this theorem, it doesn't matter whether the shape of the parent population is normal, positively skewed, negatively skewed, or some nameless, bizarre shape, as long as the sample size is sufficiently large. What constitutes "sufficiently large" depends on the shape of the parent population. If the shape of the parent population is normal, then any sample size (even a sample size of one) will be sufficiently large. Otherwise, depending on the degree of non-normality in the parent population, a sample size between 25 and 100 is sufficiently large.

1. **Why the Central Limit Theorem Works**

In a normal curve, you will recall, intermediate values are the most prevalent, and extreme values, either larger or smaller, occupy the tapered flanks. Why, when the sample size is large, does the sampling distribution approximate a normal curve, even though the parent population might be non-normal?

2. **Many Sample Means with Intermediate Values**

When the sample size is large, it is most likely that any single sample will contain the full spectrum of small, intermediate, and large scores from the parent population, whatever its shape. The calculation of a mean for this type of sample tends to neutralize or dilute the effects of any extreme scores, and the sample mean emerges with some intermediate value.

Accordingly, intermediate values prevail in the sampling distribution, and they cluster around a peak frequency representing the most common or modal value of the sample mean, as suggested at the bottom of Figure 9.3.

3. **Few Sample Means with Extreme Values**

To account for the rarer sample mean values in the tails of the sampling distribution, focus on those relatively infrequent samples that, just by chance, contain less than the full spectrum of scores from the parent population. Sometimes, because of the relatively large number of extreme scores in a particular direction, the calculation of a mean only slightly dilutes their effect, and the sample mean emerges with some more extreme value. The likelihood of obtaining extreme sample mean values declines with the extremity of the value, producing the smoothly tapered, slender tails that characterize a normal curve.

3.3 HYPOTHESIS TESTING

3.3.1 TESTING A HYPOTHESIS ABOUT SAT SCORES

Test the hypothesis that, with respect to the national average, nothing special is happening in the local population. Insofar as an investigator usually suspects just the opposite namely, that something special is happening in the local population he or she hopes to reject the hypothesis that nothing special is happening, henceforth referred to as the null hypothesis and defined more formally in a later section.

1. **Hypothesized Sampling Distribution**

If the null hypothesis is true, then the distribution of sample means that is, the sampling distribution of the mean for all possible random samples, each of size 100, from the local population of freshmen will be centered about the national average of 500. (Remember, the mean of the sampling distribution always equals the population mean).

In Figure 10.1, this sampling distribution is referred to as the hypothesized sampling distribution, since its mean equals 500, the hypothesized mean reading score for the local population of freshmen.

Anticipating the key role of the hypothesized sampling distribution in our hypothesis test, let's focus on two more properties of this distribution:

i. In Figure 10.1, vertical lines appear, at intervals of size 11, on either side of the hypothesized population mean of 500. These intervals reflect the size of the standard error of the mean, $\overline{X}$ To verify this fact, originally demonstrated in Chapter 9, substitute 110 for the population standard deviation, σ, and 100 for the sample size, n, in Formula 9.2 to obtain

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{110}{\sqrt{100}} = \frac{110}{10} = 11$$

ii. Notice that the shape of the hypothesized sampling distribution in Figure 10.1 approximates a normal curve, since the sample size of 100 is large enough to satisfy the requirements of the central limit theorem. Eventually, with the aid of normal curve tables, we will be able to construct boundaries for common and rare outcomes under the null hypothesis.

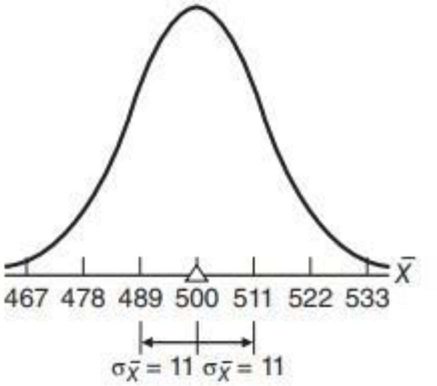


FIGURE 10.1
Hypothesized sampling distribution of the mean centered about a hypothesized population mean of 500.

The null hypothesis that the population mean for the freshman class equals 500 is tentatively assumed to be true. It is tested by determining whether the one observed sample mean qualifies as a common outcome or a rare outcome in the hypothesized sampling distribution of Figure 10.1.

2. Common Outcomes

An observed sample mean qualifies as a common outcome if the difference between its value and that of the hypothesized population mean is small enough to be viewed as a probable outcome under the null hypothesis.

That is, a sample mean qualifies as a common outcome if it doesn't deviate too far from the hypothesized population mean but appears to emerge from the dense concentration of possible sample means in the middle of the sampling distribution. A common outcome signifies a lack of evidence that, with respect to the null hypothesis, something special is happening in the underlying population. Because now there is no compelling reason for rejecting the null hypothesis, it is retained.

3. Rare Outcomes

An observed sample mean qualifies as a rare outcome if the difference between its value and the hypothesized population mean is too large to be reasonably viewed as a probable outcome under the null hypothesis.

That is, a sample mean qualifies as a rare outcome if it deviates too far from the hypothesized mean and appears to emerge from the sparse concentration of possible sample means in either tail of the sampling distribution. A rare outcome signifies that, with respect to the null hypothesis, something special probably is happening in the underlying population. Because now there are grounds for suspecting the null hypothesis, it is rejected.

4. Boundaries for Common and Rare Outcomes

Superimposed on the hypothesized sampling distribution in Figure 10.2 is one possible set of boundaries for common and rare outcomes, expressed in values of X. If the one observed sample mean is located between 478 and 522, it will qualify as a common outcome (readily attributed to variability) under the null hypothesis, and the null hypothesis will be retained. If, however, the one observed sample mean is greater than 522 or less than 478, it will qualify as a rare outcome (not readily attributed to variability) under the null hypothesis, and the null hypothesis will be rejected. Because the observed sample mean of 533 does exceed 522, the null hypothesis is rejected. On the basis of the present test, it is unlikely that the sample of 100 freshmen, with a mean math score of 533, originates from a population whose mean equals the national average of 500, and, therefore, the investigator can conclude that the mean math score for the local population of freshmen probably differs from (exceeds) the national average.

3.3.2 z TEST FOR A POPULATION MEAN

For the hypothesis test with SAT math scores, it is customary to base the test not on the hypothesized sampling distribution of X shown in Figure 10.2, but on its standardized counterpart, the hypothesized sampling distribution of z shown in Figure 10.3. Now z represents a variation on the familiar standard score, and it displays all of the properties of standard scores.

Furthermore, like the sampling distribution of , the **sampling distribution of z** represents the distribution of z values that would be obtained if a value of z were calculated for each sample mean for all possible random samples of a given size from some population.

The conversion from X to z yields a distribution that approximates the standard normal curve in Table A of Appendix C, since, as indicated in Figure 10.3, the original hypothesized population mean (500) emerges as a z score of 0 and the original standard error of the mean (11) emerges as a z score of 1. The shift from to z eliminates the original units of measurement and standardizes the hypothesis test across all situations without, however, affecting the test results.

1. Converting a Raw Score to z

To convert a raw score into a standard score, express the raw score as a distance from its mean (by subtracting the mean from the raw score), and then split this distance into standard deviation units (by dividing with the standard deviation). Expressing this definition as a word formula, we have in which, of course, the standard score indicates the deviation of the raw score in standard deviation units, above or below the mean.

2. Converting a Sample Mean to z

The z for the present situation emerges as a slight variation of this word formula: Replace the raw score with the one observed sample mean X; replace the mean with the mean of the sampling distribution, that is, the hypothesized population mean μ_{hyp};and replace the standard deviation with the standard error of the mean $\sigma_{\bar{x}}$. Now

z RATIO FOR A SINGLE POPULATION MEAN

$$z = \frac{\bar{X} - \mu_{\text{hyp}}}{\sigma_{\bar{x}}} \quad (10.1)$$

where z indicates the deviation of the observed sample mean in standard error units, above or below the hypothesized population mean.

To test the hypothesis for SAT scores, we must determine the value of z from Formula 10.1. Given a sample mean of 533, a hypothesized population mean of 500, and a standard error of 11, we find

$$z = \frac{533-500}{11} = \frac{33}{11} = 3$$

The observed z of 3 exceeds the value of 1.96 specified in the hypothesized sampling distribution in Figure 10.3. Thus, the observed z qualifies as a rare outcome under the null hypothesis, and the null hypothesis is rejected. The results of this test with z are the same as those for the original hypothesis test with $\bar{x}$.

Progress Check *10.1 Calculate the value of the z test for each of the following situations:

(a) $\bar{X} = 566; \sigma = 30; n = 36; \mu_{\text{hyp}} = 560$

(b) $\bar{X} = 24; \sigma = 4; n = 64; \mu_{\text{hyp}} = 25$

(c) $\bar{X} = 82; \sigma = 14; n = 49; \mu_{\text{hyp}} = 75$

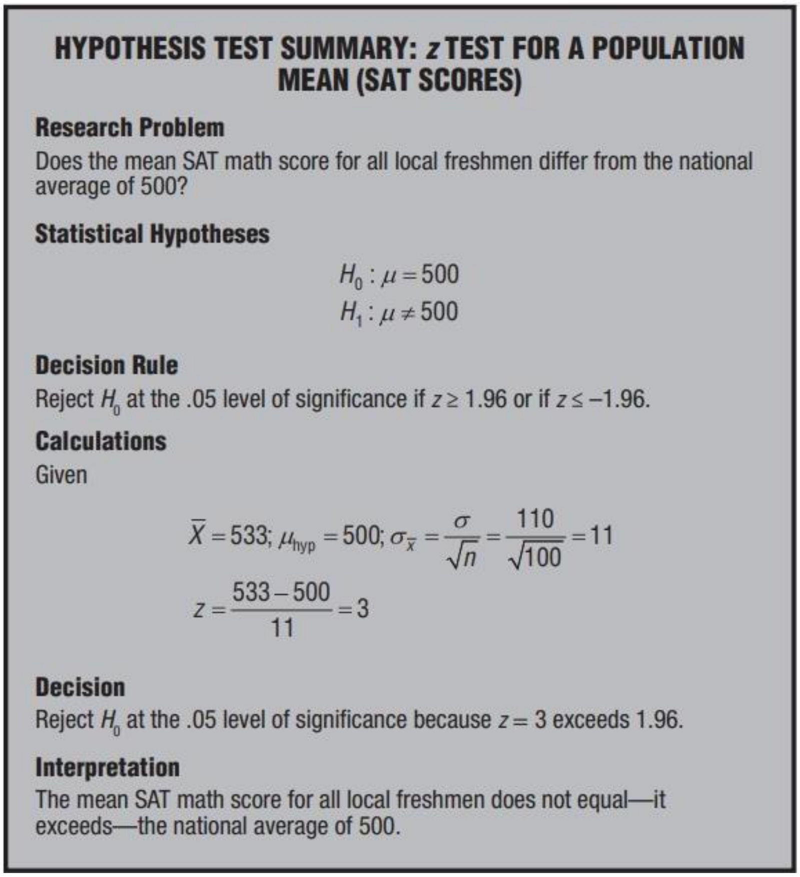
(d) $\bar{X} = 136; \sigma = 15; n = 25; \mu_{\text{hyp}} = 146$

3.3.3 STEP-BY-STEP PROCEDURE

The more important features of hypothesis testing, let’s take a detailed look at the test for SAT scores. The test procedure lends itself to a step-by-step description, beginning with a brief statement of the problem that inspired the test and ending with an interpretation of the test results. The following box summarizes the step-by-step procedure for the current hypothesis test.

3.3.4 STATEMENT OF THE RESEARCH PROBLEM

The formulation of a research problem often represents the most crucial and exciting phase of an investigation. Indeed, the mark of a skillful investigator is to focus on an important research problem that can be answered. Do children from broken families score lower on tests of personal adjustment? Do aggressive TV cartoons incite more disruptive behavior in preschool children? Does profit sharing increase the productivity of employees? Because of our emphasis on hypothesis testing, research problems appear in this book as finished products, usually in the first one or two sentences of a new example.



HYPOTHESIS TEST SUMMARY: *z* TEST FOR A POPULATION MEAN (SAT SCORES)

Research Problem

Does the mean SAT math score for all local freshmen differ from the national average of 500?

Statistical Hypotheses

$$H_0 : \mu = 500$$
$$H_1 : \mu \neq 500$$

Decision Rule

Reject H_0 at the .05 level of significance if $z \geq 1.96$ or if $z \leq -1.96$.

Calculations

Given

$$\overline{X} = 533; \mu_{\text{hyp}} = 500; \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{110}{\sqrt{100}} = 11$$

$$z = \frac{533 - 500}{11} = 3$$

Decision

Reject H_0 at the .05 level of significance because $z = 3$ exceeds 1.96.

Interpretation

The mean SAT math score for all local freshmen does not equal—it exceeds—the national average of 500.

3.3.5 NULL HYPOTHESIS (H_0)

Once the problem has been described, it must be translated into a statistical hypoth- esis regarding some population characteristic. Abbreviated as H $_0$, the null hypothesis becomes the focal point for the entire test procedure (even though we usually hope to reject it). In the test with SAT scores, the null hypothesis asserts that, with respect to the national average of 500, nothing special is happening to the mean score for the local population of freshmen. An equivalent statement, in symbols, reads:

$$H_0 : \mu = 500$$

where H_0 represents the null hypothesis and μ is the population mean for the local freshman class.

Generally speaking, the **null hypothesis (H_0)** is a statistical hypothesis that usually asserts that nothing special is happening with respect to some characteristic of the underlying population. Because the hypothesis testing procedure requires that the hypothesized sampling distribution of the mean be centered about a single number (500), the null hypothesis equals a single number (H : μ=500). Furthermore, the null hypothesis always makes a precise statement about a characteristic of the population, never about a sample. Remember, the purpose of a hypothesis test is to determine whether a particular outcome, such as an observed sample mean, could have reason- ably originated from a population with the hypothesized characteristic.

Finding the Single Number for H_0

The single number actually used in H_0 varies from problem to problem. Even for a given problem, this number could originate from any of several sources. For instance, it could be based on available information about some relevant population other than the target population, as in the present example in which 500 reflects the mean SAT math scores for all college-bound students during a recent year. It also could be based on some existing standard or theory for example, that the mean math score for the current population of local freshmen should equal 540 because that happens to be the mean score achieved by all local freshmen during recent years.

If, as sometimes happens, it's impossible to identify a meaningful null hypothesis, don't try to salvage the situation with arbitrary numbers. Instead, use another entirely different technique, known as estimation, which is described in Chapter 12.

3.3.6 ALTERNATIVE HYPOTHESIS (H_1)

In the present example, the alternative hypothesis asserts that, with respect to the national average of 500, something special is happening to the mean math score for the local population of freshmen (because the mean for the local population doesn't equal the national average of 500). An equivalent statement, in symbols, reads:

represents the alternative hypothesis, μ is the population mean for the local freshman class, and signifies, "is not equal to." The **alternative hypothesis (H_1)** asserts the opposite of the null hypothesis. A decision to retain the null hypothesis implies a lack of support for the alternative hypothesis, and a decision to reject the null hypothesis implies support for the alternative hypothesis.

3. 3.7 DECISION RULE

A **decision rule** specifies precisely when H0 should be rejected (because the observed z qualifies as a rare outcome). There are many possible decision rules, as will be seen in Section 11.3. A very common one, already introduced in Figure 10.3, specifies that H0 should be rejected if the observed z equals or is more positive than 1.96 or if the observed z equals or is more negative than –1.96. Conversely, H0 should be retained if the observed z falls between ± 1.96.

1. **Critical z Scores**

Figure 10.4 indicates that z scores of ± 1.96 define the boundaries for the middle .95 of the total area (1.00) under the hypothesized sampling distribution for z. Derived from the normal curve table, as you can verify by checking Table A in Appendix C, these two z scores separate common from rare outcomes and hence dictate whether H0 should be retained or rejected. Because of their vital role in the decision about H0 , these scores are referred to as critical z scores.

2. **Level of Significance (α)**

Figure 10.4 also indicates the proportion (.025 .025 .05) of the total area that is identified with rare outcomes. Often referred to as the level of significance of the statistical test, this proportion is symbolized by the Greek letter α (alpha). In the present example, the level of significance, α, equals .05.

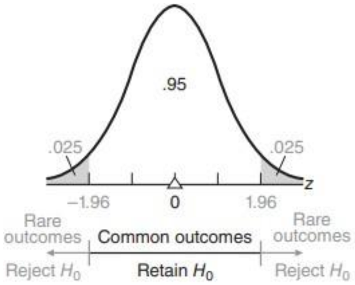


FIGURE 10.4
Proportions of area associated with common and rare outcomes (α = 05).

The level of significance (α) indicates the degree of rarity required of an observed outcome in order to reject the null hypothesis (H_0). For instance, the .05 level of significance indicates that H0 should be rejected if the observed z could have occurred just by chance with a probability of only .05 (one chance out of twenty) or less.

3.3.8 CALCULATIONS

Use information from the sample to calculate a value for z. As has been noted previously, use Formula 10.1 to convert the observed sample mean of 533 into a z of 3.

3.3.9 DECISION

Either retain or reject H_0, depending on the location, of the observed z value relative to the critical z values specified in the decision rule. According to the present rule, H_0 should be rejected at the .05 level of significance because the observed z of 3 exceeds the critical z of 1.96 and, therefore, qualifies as a rare outcome, that is, an unlikely outcome from a population centered about the null hypothesis.

1. Retain or Reject H_0?

If you are ever confused about whether to retain or reject H_0, recall the logic behind the hypothesis test. You want to reject H only if the observed value of z qualifies as a rare outcome because it deviates too far into the tails of the sampling distribution. Therefore, you want to reject H_0.

Only if the observed value of z equals or is more positive than the upper critical z (1.96) or if it equals or is more negative than the lower critical z (−1.96). Before deciding, you might find it helpful to sketch the hypothesized sampling distribution, along with its critical z values and shaded rejection regions, and then use some mark, such as an arrow (↑), to designate the location of the observed value of z (3) along the z scale. If this mark is located in the shaded rejection region or farther out than this region, as in Figure 10.4—then H_0 should be rejected.

3.3.10 INTERPRETATION

Finally, interpret the decision in terms of the original research problem. In the present example, it can be concluded that, since the null hypothesis was rejected, the mean SAT math score for the local freshman class probably differs from the national average of 500. Although not a strict consequence of the present test, a more specific conclusion is possible. Since the sample mean of 533 (or its equivalent z of 3) falls in the upper rejection region of the hypothesized sampling distribution, it can be concluded that the population mean SAT math score for all local freshmen probably exceeds the national average of 500. By the same token, if the observed sample mean or its equivalent z had fallen in the lower rejection region of the hypothesized sampling distribution, it could have been concluded that the population mean for all local freshmen probably is below the national average. If the observed sample mean or its equivalent z had fallen in the retention region of the hypothesized sampling distribution, it would have been concluded (somewhat weakly, as discussed in Section 11.2) that there is no evidence that the population mean for all local freshmen differs from the national average of 500.

3.4.1 WHY HYPOTHESIS TESTS?

There is a crucial link between hypothesis tests and the need of investigators, whether pollsters or researchers, to generalize beyond existing data. If the 100 freshmen in the SAT example of the previous chapter had been not a sample but a census of the entire freshman class, there wouldn't have been any need to generalize beyond existing data, and it would have been inappropriate to conduct a hypothesis test. Now, the observed difference between the newly observed population mean of 533 and the national average of 500, by itself, would have been sufficient grounds for concluding that the mean SAT math score for all local freshmen exceeds the national average. Indeed, any observed difference in favor of the local freshmen, regardless of the size of the difference, would have supported this conclusion.

If we must generalize beyond the 100 freshmen to a larger local population, as was actually the case, the observed difference between 533 and 500 cannot be interpreted at face value. The basic problem is that the sample mean for a second random sample of 100 freshmen probably would differ, just by chance, from the sample mean of 533 for the first sample. Accordingly, the variability among sample means must be considered when we attempt to decide whether the observed difference between 533 and 500 is real or merely transitory.

1. Importance of the Standard Error

To evaluate the effect of chance, we use the concept of a sampling distribution, that is, the concept of the sample means for all possible random outcomes. A key element in this concept is the standard error of the mean, a measure of the average amount by which sample means differ, just by chance, from the population mean. Dividing the observed difference (533−500) by the standard error (11) to obtain a value of z (3) locates the original observed difference along a z scale of either common outcomes (reasonably attributable to chance) or rare outcomes (not reasonably attributable to chance). If, when expressed as z, the ratio of the observed difference to the standard error is small enough to be reasonably attributed to chance, we retain H0. 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 H0.

Before generalizing beyond the existing data, we must always measure the effect of chance; that is, we must obtain a value for the standard error. To appreciate the vital role of the standard error in the SAT example, increase its value from 11 to 33 and note that even though the observed difference remains the same (533−500), we would retain, not reject, H0 because now z would equal 1 (rather than 3) and be less than the critical z of 1.96.

2. Possibility of Incorrect Decisions

Having made a decision about the null hypothesis, we never know absolutely whether that decision is correct or incorrect, unless, of course, we survey the entire population. Even if H0 is true (and, therefore, the hypothesized distribution of z about H0 also is true), there is a slight possibility that, just by chance, the one observed z actually originates from one of the shaded rejection regions of the hypothesized distribution of z, thus causing the true H0 to be rejected. This type of incorrect decision—rejecting a true H0—is referred to as a type I error or a false alarm.

On first impulse, it might seem desirable to abolish the shaded rejection regions in the hypothesized sampling distribution to ensure that a true H0 never is rejected. A most unfortunate consequence of this strategy, however, is that no H0, not even a radically false H0, ever would be rejected. This second type of incorrect decision—retaining a false H0—is referred to as a type II error or a miss. Both type I and type II errors are described in more detail later in this chapter.

11.2 STRONG OR WEAK DECISIONS

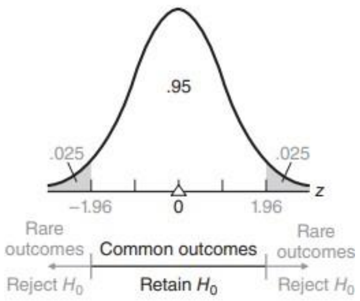


FIGURE 11.1
Proportions of area associated with common and rare outcomes (α = .05).

3. Minimizing Incorrect Decisions

Traditional hypothesis-testing procedures, such as the one illustrated in Figure 11.1, tend to minimize both types of incorrect decisions. If H_0 is true, there is a high probability that the observed z will qualify as a common outcome under the hypothesized sampling distribution and that the true H0 will be retained. (In Figure 11.1, this probability equals the proportion of white area (.95) in the hypothesized sampling distribution.)

On the other hand, if H_0 is seriously false, because the hypothesized population mean differs considerably from the true population mean, there is also a high probability that the observed z will qualify as a rare outcome under the hypothesized distribution and that the false H_0 will be rejected. (In Figure 11.1, this probability can't be determined since; in this case, the hypothesized sampling distribution does not actually reflect the true sampling distribution.

Even though we never really know whether a particular decision is correct or incorrect, it is reassuring that in the long run, most decisions will be correct— assuming the null hypotheses are either true or seriously false

3.4.2 STRONG OR WEAK DECISIONS

1. Retaining H0 Is a Weak Decision

There are subtle but important differences in the interpretation of decisions to retain H0 and to reject H_0. H_0 is retained whenever the observed z qualifies as a common outcome on the assumption that H_0 is true. Therefore, H_0 could be true. However, the same observed result also would qualify as a common outcome when the original value in H_0 (500) is replaced with a slightly different value. Thus, the retention of H_0 must be viewed as a relatively weak decision. Because of this weakness, many statisticians prefer to describe this decision as simply a failure to reject H_0 rather than as the retention of H_0. In any event, the retention of H_0 can't be interpreted as proving H_0 to be true. If H_0 had been retained in the present example, it would have been appropriate to conclude not that the mean SAT math score for all local fresh men equals the national average, but that the mean SAT math score could equal the national average, as well as many other possible values in the general vicinity of the national average.

2. Rejecting H0 Is a Strong Decision

On the other hand, H0 is rejected whenever the observed z qualifies as a rare outcome one that could have occurred just by chance with a probability of .05 or less on the assumption that H_0 is true. This suspiciously rare outcome implies that H_0 is probably false (and conversely, that H_1 is probably true). Therefore, the rejection of H_0 can be viewed as a strong decision. When H_0 was rejected in the present example, it was appropriate to report a definitive conclusion that the mean SAT math score for all local freshmen probably exceeds the national average.

To summarize,

The decision to retain H_0 implies not that H_0 is probably true, but only that H0 could be true, whereas the decision to reject H_0 implies that H_0 is probably false (and that H_1 is probably true).

Since most investigators hope to reject H_0 in favor of H_1, the relative weakness of the decision to retain H_0 usually does not pose a serious problem.

3. Why the Research Hypothesis Isn't Tested Directly

Even though H0, the null hypothesis, is the focus of a statistical test, it is usually of secondary concern to the investigator. Nevertheless, there are several reasons why, although of primary concern, the research hypothesis is identified with H1 and tested indirectly.

4. Lacks Necessary Precision

The research hypothesis, but not the null hypothesis, lacks the necessary precision to be tested directly.

To be tested, a hypothesis must specify a single number about which the hypothesized sampling distribution can be constructed. Because it specifies a single number, the null hypothesis, rather than the research hypothesis, is tested directly. In the SAT example, the null hypothesis specifies that a precise value (the national average of 500) describes the mean for the current population of interest (all local freshmen). Typically, the research hypothesis lacks the required precision. It merely specifies that some inequality exists between the hypothesized value (500) and the mean for the current population of interest (all local freshmen).

5. Supported by a Strong Decision to Reject

Logical considerations also argue for the indirect testing of the research hypothesis and the direct testing of the null hypothesis.

Because the research hypothesis is identified with the alternative hypothesis, the decision to reject the null hypothesis, should it be made, will provide strong support for the research hypothesis, while the decision to retain the null hypothesis, should it be made, will provide, at most, weak support for the null hypothesis.

As mentioned, the decision to reject the null hypothesis is stronger than the decision to retain it. Logically, a statement such as "All cows have four legs" can never be proven in spite of a steady stream of positive instances. It only takes one negative instance—one cow with three legs—to disprove the statement. By the same token, one positive instance (common outcome) doesn't prove the null hypothesis, but one negative instance (rare outcome) disproves the null hypothesis. (Strictly speaking, however, since a rare outcome implies that the null hypothesis is probably but not definitely false, remember that there always is a very small possibility that the rare outcome reflects a true null hypothesis).

Logically, therefore, it makes sense to identify the research hypothesis with the alternative hypothesis. If, as hoped, the data favor the research hypothesis, the test will generate strong support for your hunch: It's probably true. If the data do not favor the research hypothesis, the hypothesis test will generate, at most, weak support for the null hypothesis: It could be true. Weak support for the null hypothesis is of little consequence, as this hypothesis that nothing special is happening in the population usually serves only as a convenient testing device.

3.4.3 ONE-TAILED AND TWO-TAILED TESTS

Two-Tailed Test Generally, the alternative hypothesis, H1, is the complement of the null hypothesis, H0. Under typical conditions, the form of H1 resembles that shown for the SAT example, namely,

This alternative hypothesis says that the null hypothesis should be rejected if the mean reading score for the population of local freshmen differs in either direction from the national average of 500. An observed z will qualify as a rare outcome if it deviates too far either below or above the national average. Panel A of Figure 11.2 shows rejection regions that are associated with both tails of the hypothesized sampling distribution. The corresponding decision rule, with its pair of critical z scores of ±1.96, is referred to as a two-tailed or nondirectional test.

1. **One-Tailed Test (Lower Tail Critical)**

Now let's assume that the research hypothesis for the investigation of SAT math scores was based on complaints from instructors about the poor preparation of local freshmen. Assume also that if the investigation supports these complaints, a remedial program will be instituted. Under these circumstances, the investigator might prefer a hypothesis test that is specially designed to detect only whether the population mean math score for all local freshmen is less than the national average.

This alternative hypothesis reads:

It reflects a concern that the null hypothesis should be rejected only if the population mean math score for all local freshmen is less than the national average of 500. Accord ingly, an observed z triggers the decision to reject H0 only if z deviates too far below the national average. Panel B of Figure 11.2 illustrates a rejection region that is associated with only the lower tail of the hypothesized sampling distribution. The corresponding decision rule, with its critical z of −1.65, is referred to as a one-tailed or directional test with the lower tail critical. Use Table A in Appendix C to verify that if the critical z equals −1.65; then .05 of the total area under the distribution of z has been allocated to the lower rejection region. Notice that the level of significance, α, equals .05 for this one-tailed test and also for the original two-tailed test.

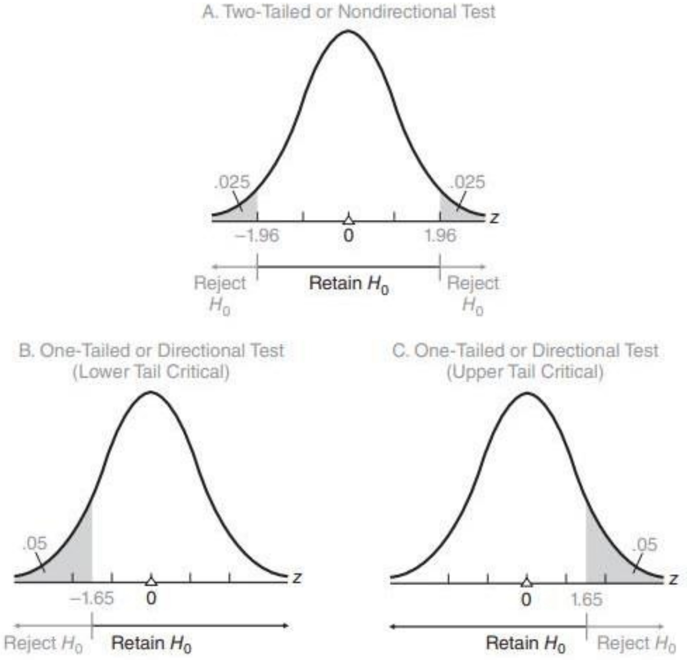

A. Two-Tailed or Nondirectional Test
.025
.025
z
−1.96
0
1.96
Reject H_0
Retain H_0
Reject H_0
B. One-Tailed or Directional Test (Lower Tail Critical)
.05
z
−1.65
0
Reject H_0
Retain H_0
C. One-Tailed or Directional Test (Upper Tail Critical)
.05
z
0
1.65
Retain H_0
Reject H_0

FIGURE 11.2
Three different types of tests ($\alpha = .05$).

2. Extra Sensitivity of One-Tailed Tests

This new one-tailed test is extra sensitive to any drop in the population mean for the local freshmen below the national average. If H0 is false because a drop has occurred, then the observed z will be more likely to deviate below the national average. As can be seen in panels A and B of Figure 11.2, an observed deviation in the direction of concern below the national average is more likely to penetrate the broader rejection region for the one-tailed test than that for the two-tailed test. Therefore, the decision to reject a false H0 (in favor of the research hypothesis) is more likely to occur in the one-tailed test than in the two-tailed test.

3. One-Tailed Test (Upper Tail Critical)

Panel C of Figure 11.2 illustrates a one-tailed or directional test with the upper tail critical. This one-tailed test is the mirror image of the previous test. Now the alternative hypothesis reads:

$$H_1 : \mu > 500$$

and its critical z equals 1.65. This test is specially designed to detect only whether the population mean math score for all local freshmen exceeds the national average. For example, the research hypothesis for this investigation might have been inspired by the possibility of eliminating an existing remedial math program if it can be demonstrated that, on the average, the SAT math scores of all local freshmen exceed the national average.

4. One or Two Tails?

Before a hypothesis test, if there is a concern that the true population mean differs from the hypothesized population mean only in a particular direction, use the appropriate one-tailed or directional test for extra sensitivity. Otherwise, use the more customary two-tailed or nondirectional test. Having committed yourself to a one-tailed test with its single rejection region, you must retain H_0, regardless of how far the observed z deviates from the hypothesized population mean in the direction of "no concern." For instance, if a one-tailed test with the lower tail critical had been used with the data for 100 freshmen from the SAT example, H_0 would have been retained because, even though the observed z equals an impressive value of 3, it deviates in the direction of no concern in this case, above the national average. Clearly, a one-tailed test should be adopted only when there is absolutely no concern about deviations, even very large deviations, in one direction. If there is the slightest concern about these deviations, use a two-tailed test. The selection of a one- or two-tailed test should be made before the data are collected. Never "peek" at the value of the observed z to determine whether to locate the rejection region for a one-tailed test in the upper or the lower tail of the distribution of z. To qualify as a one-tailed test, the location of the rejection region must reflect the investigator's concern only about deviations in a particular direction before any inspection of the data. Indeed, the investigator should be able to muster a compelling reason, based on an understanding of the research hypothesis, to support the direction of the one-tailed test.

5. New Null Hypothesis for One-Tailed Tests

When tests are one-tailed, a complete statement of the null hypothesis also should include all possible values of the population mean in the direction of no concern. For example, given a one-tailed test with the lower tail critical, such as H1: $\mu < 500$, the complete null hypothesis should be stated as H0: $\mu \geq 500$ instead of H0: $\mu = 500$. By the same token, given a one-tailed test with the upper tail critical, such as H1: $\mu > 500$, the complete null hypothesis should be stated as H0: $\mu \leq 500$. If you think about it, the complete H0 describes all of the population means that could be true if a one-tailed test results in the retention of the null hypothesis. For instance, if a one-tailed test with the lower tail critical results in the retention of H0: $\mu \geq 500$, the complete H0 accurately reflects the fact that not only $\mu = 500$ could be true, but also that any other value of the population mean in the direction of no concern, that is, $\mu > 500$, could be true. (Remember, when the test is one-tailed, even a very deviant result in the direction of no concern possibly reflecting a mean much larger than 500 still would trigger the decision to retain H0.) Henceforth, whenever a one-tailed test is employed, write H0 to include values of the population mean in the direction of no concern even though the single number in the complete H0 identified by the equality sign is the one value about which the hypothesized sampling distribution is centered and, therefore, the one value actually used in the hypothesis test.

3.5 ESTIMATION

3.5.1 POINT ESTIMATE FOR μ

A point estimate for μ uses a single value to represent the unknown population mean. This is the most straightforward type of estimate. If a random sample of 100 local freshmen reveals a sample mean SAT score of 533, then 533 will be the point estimate of the unknown population mean for all local freshmen. The best single point estimate for the unknown population mean is simply the observed value of the sample mean.

A Basic Deficiency

Although straightforward, simple, and precise, point estimates suffer from a basic deficiency. They tend to be inaccurate. Because of sampling variability, it's unlikely that a single sample mean, such as 533, will coincide with the population mean. Since point estimates convey no information about the degree of inaccuracy due to sampling variability, statisticians supplement point estimates with another, more realistic type of estimate, known as interval estimates or confidence intervals.

3.5.2 CONFIDENCE INTERVAL (CI) FOR μ

A confidence interval for μ uses a range of values that, with a known degree of certainty, includes the unknown population mean. For instance, the SAT investigator might use a confidence interval to claim, with 95 percent confidence, that the interval between 511.44 and 554.56 includes the population mean math score for all local freshmen. To be 95 percent confident signifies that if many of these intervals were constructed for a long series of samples, approximately 95 percent would include the population mean for all local freshmen. In the long run, 95 percent of these confidence intervals are true because they include the unknown population mean. The remaining 5 percent are false because they fail to include the unknown population mean.

1. Why Confidence Intervals Work

To understand confidence intervals, you must view them in the context of three important properties of the sampling distribution of the mean.

For the sampling distribution from which the sample mean of 533 originates, as shown in Figure 12.1, the three important properties are as follows:

- The mean of the sampling distribution equals the unknown population mean for all local freshmen, whatever its value, because the mean of this sampling distribution always equals the population mean.
- The standard error of the sampling distribution equals the value (11) obtained from dividing the population standard deviation (110) by the square root of the sample size $(\sqrt{100})$
- The shape of the sampling distribution approximates a normal distribution because the sample size of 100 satisfies the requirements of the central limit theorem.

2. A Series of Confidence Intervals

In practice, only one sample mean is actually taken from this sampling distribution and used to construct a single 95 percent confidence interval. However, imagine taking not just one but a series of randomly selected sample means from this sampling distribution. Because of sampling variability, these sample means tend to differ among themselves. For each sample mean, construct a 95 percent confidence interval by adding 1.96 standard errors to the sample mean and subtracting 1.96 standard errors from the sample mean; that is, use the expression.

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

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

3. True Confidence Intervals

According to statistical theory, do 95 percent of these confidence intervals include the unknown population mean? As indicated in Figure 12.2, because the sampling distribution is normal, 95 percent of all sample means are within 1.96 standard errors of the unknown population mean, that is, 95 percent of all sample means deviate less than 1.96 standard errors from the unknown population mean. Therefore, and this is the key point, when sample means are expanded into confidence intervals by adding and subtracting 1.96 standard errors 95 percent of all possible confidence intervals are true because they include the unknown population mean. To illustrate

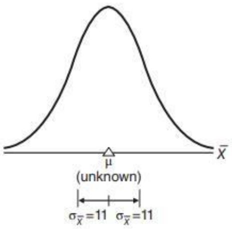


FIGURE 12.1
Sampling distribution of the mean (SAT scores).

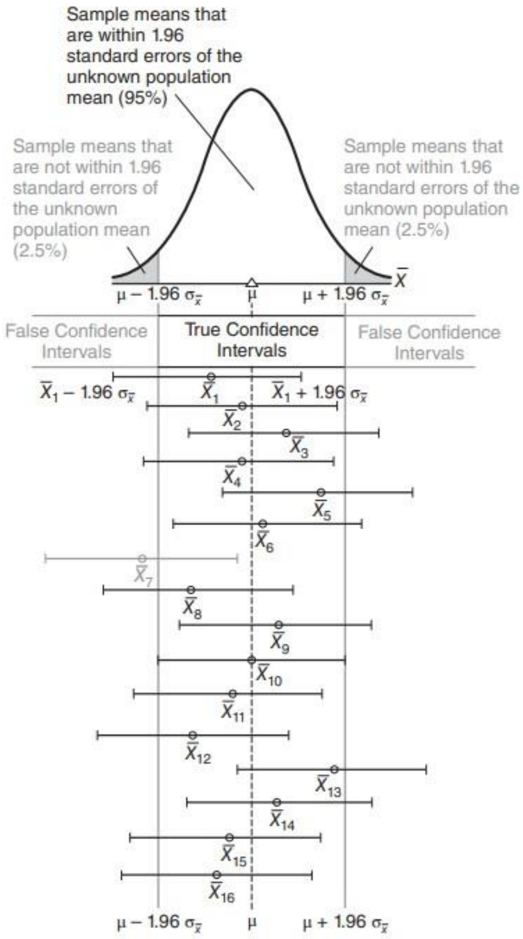

Sample means that are within 1.96 standard errors of the unknown population mean (95%)
Sample means that are not within 1.96 standard errors of the unknown population mean (2.5%)
Sample means that are not within 1.96 standard errors of the unknown population mean (2.5%)
$\overline{X}$
$\mu - 1.96\,\sigma_{\overline{X}}$
μ
$\mu + 1.96\,\sigma_{\overline{X}}$
False Confidence Intervals
True Confidence Intervals
False Confidence Intervals
$\overline{X}_1 - 1.96\,\sigma_{\overline{X}}$
$\overline{X}_1$
$\overline{X}_1 + 1.96\,\sigma_{\overline{X}}$
$\overline{X}_2$
$\overline{X}_3$
$\overline{X}_4$
$\overline{X}_5$
$\overline{X}_6$
$\overline{X}_7$
$\overline{X}_8$
$\overline{X}_9$
$\overline{X}_{10}$
$\overline{X}_{11}$
$\overline{X}_{12}$
$\overline{X}_{13}$
$\overline{X}_{14}$
$\overline{X}_{15}$
$\overline{X}_{16}$
$\mu - 1.96\,\sigma_{\overline{X}}$
μ
$\mu + 1.96\,\sigma_{\overline{X}}$

FIGURE 12.2

A series of 95 percent confidence intervals (emerging from a sampling distribution).

this point, 15 of the 16 sample means shown in Figure 12.2 are within 1.96 standard errors of the unknown population mean. The corresponding 15 confidence intervals have ranges that span the broken line for the population mean, thereby qualifying as true intervals because they include the value of the unknown population mean.

4. False Confidence Intervals

Five percent of all confidence intervals fail to include the unknown population mean. As indicated in Figure 12.2, 5 percent of all sample means (2.5 percent in each tail) deviate more than 1.96 standard errors from the unknown population mean. Therefore, when sample means are expanded into confidence intervals—by adding and subtracting 1.96 standard errors—5 percent of all possible confidence intervals are false because they fail to include the unknown population mean. To illustrate this point, only 1 of the 16 sample means shown in Figure 12.2 is not within 1.96 standard errors of the unknown population mean.

The resulting confidence interval, shown as shaded, has a range that does not span the broken line for the population mean, thereby being designated as a false interval because it fails to include the value of the unknown population mean.

5. Confidence Interval for μ Based on z

To determine the previously reported confidence interval of 511.44 to 554.56 for the unknown mean math score of all local freshmen, use the following general expression:

CONFIDENCE INTERVAL FOR μ (BASED ON z)

$$\overline{X} \pm (z_{conf})(\sigma_{\overline{X}}) \qquad (12.1)$$

where $\overline{X}$ presents the sample mean; z_{conf} represents a number from the standard normal table that satisfies the confidence specifications for the confidence interval; and σx represents the standard error of the mean. Given that $\overline{X}$, the sample mean SAT math score, equals 533, that z_{conf} equals 1.96 (from the standard normal tables, where z scores of ±1.96 define the middle 95 percent of the area under the normal curve), and that the standard error, σ_x, equals 11, Formula 12.1 becomes wher $\overline{X}$ represents the sample mean; z_{conf} represents a number from the standard normal table that satisfies the confidence specifications for the confidence interval; and σx represents the standard error of the mean. Given that $\overline{X}$ he sample mean SAT math score, equals 533, that z_{conf} equals 1.96 (from the standard normal tables, where z scores of ±1.96 define the middle 95 percent of the area under the normal curve), and that the standard error, σx, equals 11, Formula 12.1 becomes

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

where 554.56 and 511.44 represent the upper and lower limits of the confidence inter val. Now it can be claimed, with 95 percent confidence, that the interval between 511.44 and 554.56 includes the value of the unknown mean math score for all local freshmen.

3.5.3 INTERPRETATION OF A CONFIDENCE INTERVAL

A 95 percent confidence claim reflects a long-term performance rating for an extended series of confidence intervals. If a series of confidence intervals is constructed to estimate the same population mean, as in Figure 12.2, approximately 95 percent of these intervals should include the population mean. In practice, only one confidence interval, not a series of intervals, is constructed, and that one interval is either true or false, because it either includes the population mean or fails to include the population mean. Of course, we never really know whether a particular confidence interval is true or false unless the entire population is surveyed. However, when the level of confidence equals 95 percent or more, we can be reasonably confident that the one observed confidence interval includes the true population mean.

For instance, we can be reasonably confident that the true population mean math score for all local freshmen is neither less than 511.44 nor more than 554.56. That's the same as being reasonably confident that the true population mean for all local freshmen is between 511.44 and 554.56.

3.5.4 LEVEL OF CONFIDENCE

The level of confidence indicates the percent of time that a series of confidence intervals includes the unknown population characteristic, such as the population mean. Any level of confidence may be assigned to a confidence interval merely by substituting an appropriate value for z_{conf} in Formula 12.1. For instance, to construct a 99 percent confidence interval from the data for SAT math scores, first consult Table A in Appendix C to verify that z_{conf} values of ±2.58 define the middle 99 percent of the total area under the normal curve. Then substitute numbers for symbols in Formula 12.1 to obtain

It can be claimed, with 99 percent confidence, that the interval between 504.62 and 561.38 includes the value of the unknown mean math score for all local freshmen. This implies that, in the long run, 99 percent of these confidence intervals will include the unknown population mean.

1. **Effect on Width of Interval**

Notice that the 99 percent confidence interval of 504.62 to 561.38 is wider and, therefore, less precise than the corresponding 95 percent confidence interval of 511.44 to 554.56. The shift from a 95 percent to a 99 percent level of confidence requires an increase in the value of zconf from 1.96 to 2.58. This increase, in turn, causes a wider, less precise confidence interval. Any shift to a higher level of confidence always produces a wider, less precise confidence interval unless offset by an increase in sample size.

2. **Choosing a Level of Confidence**

Although many different levels of confidence have been used, 95 percent and 99 percent are the most prevalent. Generally, a larger level of confidence, such as 99 per cent, should be reserved for situations in which a false interval might have particularly serious consequences, such as the failure of a national opinion pollster to predict the winner of a presidential election.

3.5.5 EFFECT OF SAMPLE SIZE

The larger the sample size, the smaller the standard error and, hence, the more precise (narrower) the confidence interval will be. Indeed, as the sample size grows larger, the standard error will approach zero and the confidence interval will shrink to a point estimate. Given this perspective, the sample size for a confidence interval, unlike that for a hypothesis test, never can be too large.

Selection of Sample Size

The hypothesis tests, sample size can be selected according to specifications established before the investigation. To generate a confidence interval that possesses the desired precision (width). Valid use of these formulas requires that before the investigation, the population standard deviation be either known or estimated.

UNIT III

INFERENTIAL STATISTICS

Populations – samples – random sampling – Sampling distribution- standard error of the mean - Hypothesis testing – z-test – z-test procedure –decision rule – calculations – decisions – interpretations - one-tailed and two-tailed tests – Estimation – point estimate – confidence interval – level of confidence – effect of sample size.

PART - A

1) What is population?

In statistics, population is the entire set of items from which you draw data for a statistical study. It can be a group of individuals, a set of items, etc. It makes up the data pool for a study.

2) What is a sample?

A sample represents the group of interest from the population, which you will use to represent the data. The sample is an unbiased subset of the population that best represents the whole data.

3) When are samples used?

- The population is too large to collect data.
- The data collected is not reliable.
- The population is hypothetical and is unlimited in size. Take the example of a study thatdocuments the results of a new medical procedure. It is unknown how the procedure willaffect people across the globe, so a test group is used to find out how people react to it.

4) Difference Between Population and Sample?

Population	Samples
All residents of a country would constitute the Population set	All residents who live above the poverty line would be the Sample
All residents above the poverty line in a	All residents who are millionaires would make

country would be the Population	up the Sample
All employees in an office would be the Population	Out of all the employees, all managers in the office would be the Sample

5) **Define Hypothetical Population**

A population containing a finite number of individuals, members or units is a class. ... All the 400 students of 10th class of particular school is an example of existent type of population and the population of heads and tails obtained by tossing a coin on infinite number of times is an example of hypothetical population.

6) **What Is Random Samplings**

Random sampling occurs if, at each stage of sampling, the selection process guarantees that all potential observations in the population have an equal chance of being included in the sample

8) What is Sampling Distribution ?

The sampling distribution of the mean refers to the probability distribution of means for all possible random samples of a given size from some population.

9) What are the types of Sampling Distribution?

- Sampling distribution of mean
- Sampling distribution of proportion
- T-distribution

10) Define Sampling distribution of mean

The most common type of sampling distribution is of the mean. It focuses on calculating the mean of every sample group chosen from the population and plotting the data points. The graph shows a normal distribution where the center is the mean of the sampling distribution, which represents the mean of the entire population.

11) What is mean by Sampling distribution of proportion

This sampling distribution focuses on proportions in a population. Samples are selected and their proportions are calculated. The mean of the sample proportions from each group represent the proportion of the entire population,

12) Define T-distribution

A T-distribution is a sampling distribution that involves a small population or one where not much is known about it. It is used to estimate the mean of the population and other statistics such as confidence intervals, statistical differences and linear regression. The T-distribution uses a t- score to evaluate data that wouldn't be appropriate for a normal distribution.

The formula for t-score is: $\boxed{\mathbf{t = [\ x - \mu\ ] / [\ s / sqrt(\ n\ )\ ]}}$

In the formula, "x" is the sample mean and "μ" is the population mean and signifies standard deviation.

13) Define MEAN OF ALL THE SAMPLE MEAN

The mean of the sampling distribution of the mean always equals the mean of the population.

14) Standard Error Of The Mean

The standard error of the mean equals the standard deviation of the population divided by the square root of the sample size

15) What is the Special Type Of Standard Deviation

You might find it helpful to think of the standard error of the mean as a rough measure of the average amount by which sample means deviate from the mean of the sampling distribution or from the population mean.

16) What Is The Hypothesis Testing

Hypothesis testing is a form of statistical inference that uses data from a sample to drawconclusions about a population parameter or a population probability distribution. First, a tentative assumption is made about the parameter or distribution. This assumption is called the null hypothesis and is denoted by H0.

17) Hypothesized Sampling Distribution

When you perform a hypothesis test of a single population mean μ using a normal distribution (often called a z-test), you take a simple random sample from the population. ....Then the binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with μ = p and σ=√pqn σ = p q n .

18) Define Decision Rule

A decision rule specifies precisely when H0 should be rejected (because the observed z qualifies as a rare outcome). There are many possible decision rules, as will be seen in Section 11.3. A very common one, already introduced in Figure 10.3, specifies that H0 should be rejected if the observed z equals or is more positive than 1.96 or if the observed z equals or is more negative than –1.96. Conversely, H0 should be retained if the observed z falls between ± 1.96.

19) Define null hypothesis?

The null hypothesis is a typical statistical theory which suggests that no statistical relationship and significance exists in a set of given single observed variable, between two sets of observed data and measured phenomena.

20) What is Level of Significance

Total area that is identified with rare outcomes. Often referred to as the level of significance ofthe statistical test, this proportion is symbolized by the Greek letter α (alpha) and discussed more thoroughly in Section 11.4. In the present example, the level of significance, α, equals 05.

21) Define One-Tailed And Two-Tailed Tests

Before a hypothesis test, if there is a concern that the true population mean differs from the hypothesized population mean only in a particular direction, use the appropriate one-tailed or directional test for extra sensitivity. Otherwise, use the more customary two-tailed or non directional test

22) What is Two-Tailed Test with example

Generally, the alternative hypothesis, H1, is the complement of the null hypothesis, H0. Under typical conditions, the form of H1 resembles that shown for the SAT example, namely,

$$H1: \mu \neq 500$$

This alternative hypothesis says that the null hypothesis should be rejected if the mean reading score for the population of local freshmen differs in either direction from the national average of 500. An observed z will qualify as a rare outcome if it deviates too far either below or above the national average. Panel A of Figure 11.2 shows rejection regions that are associated with both tails of the hypothesized sampling distribution. The corresponding decision rule, with its pair of critical z scores of ±1.96, is referred to as a two-tailed or non directional test.

23) what is One-Tailed Test (Lower Tail Critical)

Now let’s assume that the research hypothesis for the investigation of SAT math scores was based on complaints from instructors about the poor preparation of local freshmen. Assume also that if the investigation supports these complaints, a remedial program will be instituted. Under these circumstances, the investigator might prefer a hypothesis test that is specially designed to detect only whether the population mean math score for all local freshmen is less than the national average. This alternative hypothesis reads:

$$H1: \mu \leq 500$$

24) What is One-Tailed Test (Upper Tail Critical)

Panel C of Figure 11.2 illustrates a one-tailed or directional test with the upper tail critical. This one-tailed test is the mirror image of the previous test. Now the alternative hypothesis reads:

$$H1: \mu > 500$$

and its critical z equals 1.65. This test is specially designed to detect only whether the populationmean math score for all local freshmen exceeds the national average. For example, the research hypothesis for this investigation might have been inspired by the possibility of eliminating an existing remedial math program if it can be demonstrated that, on the average, the SAT math scores of all local freshmen exceed the national average

25) Define Consequences of Reducing Standard Error

As can be seen by comparing Figure 11.5 and Figure 11.6, the reduction of the standard error from 2.5 to 1.5 has two important consequences:

1. It shrinks the upper retention region back toward the hypothesized population mean of 100.
2. It shrinks the entire true sampling distribution toward the true population mean of 103.

26) Define Power curve

A graph showing power as a function of some other variable; specifically a graph of the power output of a vehicle or aircraft against engine speed. 2 figurative Chiefly Business. The current thinking or trend. 3Statistics. A graphical representation of the power function of a statistical test.

27) For a one-tailed or directional test with the lower tail critical

$$H0: \mu \geq \text{SOME NUMBERS}$$

$$H1: \mu < \text{SOME NUMBERS}$$

28) For a one-tailed or directional test with the upper tail critical,

$$H0: \mu \leq \text{SOME NUMBERS}$$

$$H1: \mu > \text{SOME NUMBERS}$$

29) What are four possible outcomes for any hypothesis test:

- If H0 really is true, it is a correct decision to retain the true H0.

- If H0 really is true, it is a type I error to reject the true H0.

- If H0 really is false, it is a type II error to retain the false H0.

- If H0 really is false, it is a correct decision to reject the false H0.

30) Define Point Estimate

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

31) What is mean by confidence interval (ci) for μ

A confidence interval for μ uses a range of values that, with a known degree of certainty, includes the unknown population mean.

32) Define Effect Of Sample Size

The larger the sample size, the smaller the standard error and, hence, the more precise (narrower) the confidence interval will be. Indeed, as the sample size grows larger, the standard error will approach zero and the confidence interval will shrink to a point estimate. Given this perspective, the sample size for a confidence interval, unlike that for a hypothesis test, never can be too large.

PART B

1) Explain population and samples. And difference?

2) Describe random sampling?

3) Explain sampling distribution and types?

4) Describe null hypothesis test in detail?

5) Explain in detail hypothesis testing and examples?

6) Does the mean of SAT math score for all local freshman differ for all local average of 500? (ztest for population mean)

7) Explain one tailed and two tailed test.

8) Define estimation .Explain in detail about point estimation.

AD3491 FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

UNIT IV Notes

1) Define T-Test?

Statistical method for the comparison of the mean of the two groups of the normally distributed sample(s).

It is used when:

- Population parameter (mean and standard deviation) is not known
- Sample size (number of observations) < 30

2) T-Test: (Explanation)

Type of t-test.

The T-test is mainly classified into 3 parts:

- One sample
- Independent sample
- Paired sample

1. One Sample

In one sample t-test, we compare the sample mean with the population mean.

Mathematical Formula:

$$t = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$\overline{X} = \textit{Sample mean}$$

$$\mu = \textit{Population mean}$$

$$\sigma = \textit{sample standard deviation}$$

$$n = \textit{sample size}$$

- Region of rejection lies either on extreme left or extreme right of the distribution.
- In z-test, we use population standard deviation instead of sample standard deviation.

Example:

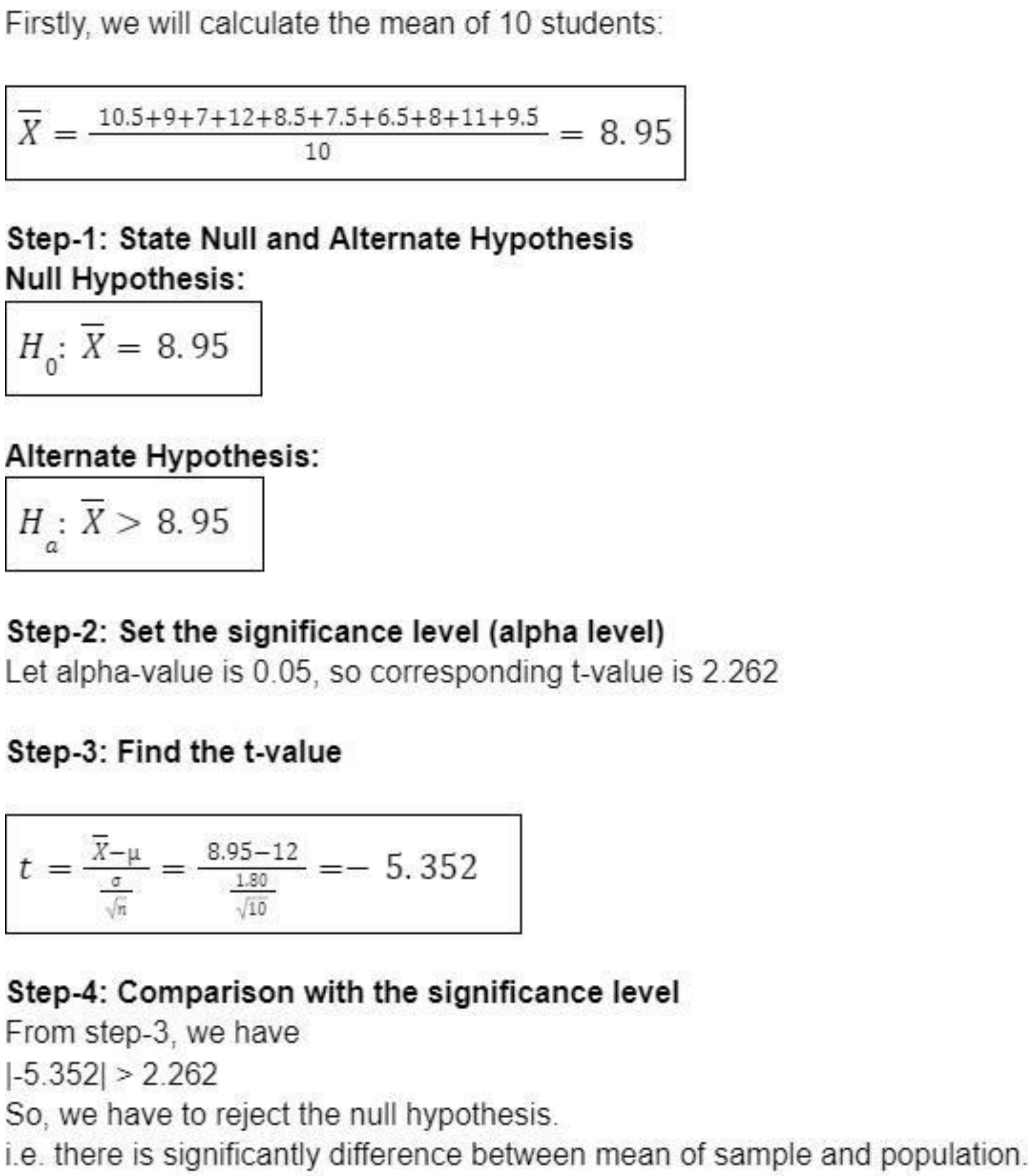
Problem Statement:

Marks of student are 10.5, 9, 7, 12, 8.5, 7.5, 6.5, 8, 11 and 9.5.

Mean population score is 12 and standard deviation is 1.80.

Is the mean value for student significantly differ from the mean population value.

Solution:



Firstly, we will calculate the mean of 10 students:

$$\boxed{\overline{X} = \frac{10.5+9+7+12+8.5+7.5+6.5+8+11+9.5}{10} = 8.95}$$

Step-1: State Null and Alternate Hypothesis
Null Hypothesis:

$$\boxed{H_0: \overline{X} = 8.95}$$

Alternate Hypothesis:

$$\boxed{H_a: \overline{X} > 8.95}$$

Step-2: Set the significance level (alpha level)
Let alpha-value is 0.05, so corresponding t-value is 2.262

Step-3: Find the t-value

$$\boxed{t = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{8.95-12}{\frac{1.80}{\sqrt{10}}} =- 5.352}$$

Step-4: Comparison with the significance level
From step-3, we have
|-5.352| > 2.262
So, we have to reject the null hypothesis.
i.e. there is significantly difference between mean of sample and population.

2. **Independent (two-sample t-test):**

In this test, we compare the means of two different samples.
Mathematical Formula:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\overline{X}_1, \overline{X}_2$: *Sample Mean*

n_1, n_2: *Sample Size*

s^2: *estimator of common variance such that*

$$s^2 = \frac{\Sigma(x-\overline{X}_1)^2 + \Sigma(x-\overline{X}_2)^2}{(n_1-1)+(n_2-1)}, \text{ where}$$

$(n_1 - 1) + (n_2 - 1)$: degree of freedom

Degree of Freedom:

Degree of freedom is defined as the number of independent variables.
It is given by:

$$df = \Sigma(n_i - 1),$$

Where

df = *degree of freedom*

n_i = *sample size*

Note: There are two regions of rejection, one in either directions towards tail of each distribution.

Let’s understand two-sample t-test by an example:

Problem Statement:

The marks of boys and girls are given:

Boys: 12, 14, 10, 8, 16, 5, 3, 9, and 11

Girls: 21, 18, 14, 20, 11, 19, 8, 12, 13, and 15

Is there any significant differnece between marks of males and females i.e. population means are different.

Solution:

Firstly, we will calculate mean, standard deviation and degree of freedom for marks of boys and girls:
Boys:

$$\boxed{\begin{array}{l} N_1 = 9, \\ df = (9-1) = 8 \\ \overline{X}_1 = 9.778,\ s_1 = 4.1164 \end{array}}$$

Girls:

$$\boxed{\begin{array}{l} N_2 = 10, \\ df = (10-1) = 9 \\ \overline{X}_2 = 15.1,\ s_2 = 4.2805 \end{array}}$$

Step-1: State Null and Alternate Hypothesis:

Null Hypothesis:

$$\boxed{H_0\text{: } \mu_1 = \mu_2}$$

Alternate Hypothesis:

$$\boxed{H_0\text{: } \mu_1 \neq \mu_2}$$

Step-2: Set the significance level (alpha level)
Let the alpha-value is 0.05, and
since the degree of freedom is 9+8=17
So, t-value is 2.11

Step-3: Find the t-value

$$\boxed{t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}} = \frac{9.778-15.1}{\sqrt{\frac{(4.1164)^2}{9} + \frac{(4.2805)^2}{10}}} = \frac{-5.322}{1.93} = -\ 2.758}$$

Step-4: Comparison with the significance level
From step-3, we have
|-2.758| > 2.11
So, we have to reject the null hypothesis.
i.e. population means are different

Firstly, we will calculate mean, standard deviation and degree of freedom for marks of boys and girls:

3. **Paired t-test:**

In this test, we compare the means of two related or same group at two different time.

Mathematical Formula:

$$t = \frac{m}{\frac{s}{\sqrt{n}}}$$

$m\text{: } mean\ of\ difference\ between\ each\ pair\ of\ values$

$s\text{: } standard\ deviation\ of\ difference\ between\ each\ pair\ of\ values$

$n\text{: } sample\ size$

Note: Degree of freedom is n-1.

Let’s understand two-sample t-test by an example:

Problem Statement:

Blood pressure of 8 patients are before and after are
recorded:Before: 180, 200, 230, 240, 170, 190, 200, and 165
After: 140, 145, 150, 155, 120, 130, 140, and 130
Is there any significant difference between BP reading before and

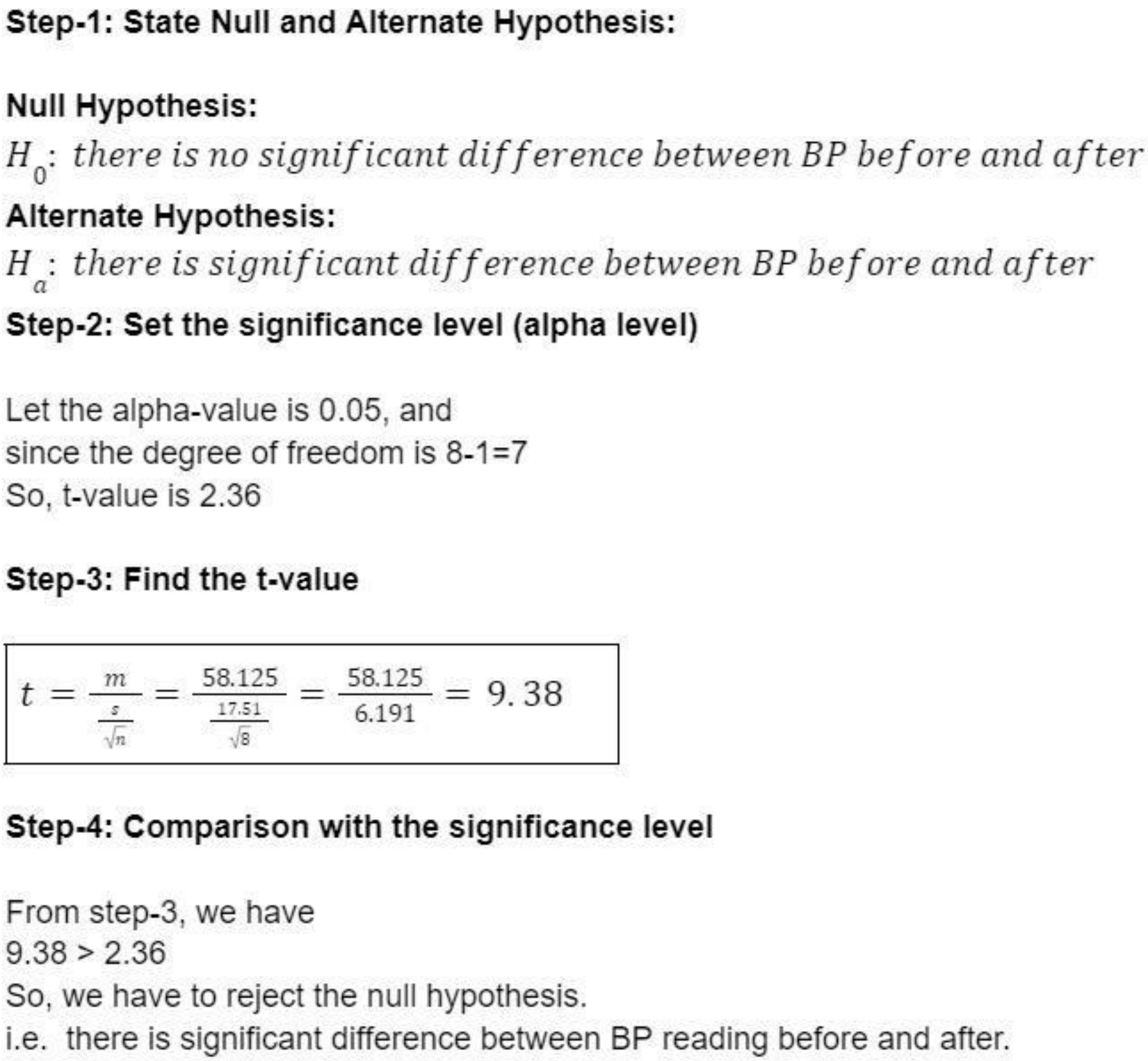
after.Solution:

Firstly, we will find the mean and standard deviation of difference between each pair of values

Before	After	d (= Before - After)	d^2
180	140	40	1600
200	145	55	3025
230	150	80	6400
240	155	85	7225
170	120	50	2500
190	130	60	3600
200	140	60	3600
165	130	35	1225
		$\Sigma d \ = \ 465$	$\Sigma d^2 = 29175$

$$\boxed{Mean\ (m) \ = \ \frac{\Sigma d}{8} = \frac{465}{8} = 58.125}$$

$$\boxed{s \ = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}} = \sqrt{\frac{(29175) - \frac{(465)^2}{8}}{8-1}} = 17.51}$$



Step-1: State Null and Alternate Hypothesis:

Null Hypothesis:

H_0: *there is no significant difference between BP before and after*

Alternate Hypothesis:

H_a: *there is significant difference between BP before and after*

Step-2: Set the significance level (alpha level)

Let the alpha-value is 0.05, and
since the degree of freedom is 8-1=7
So, t-value is 2.36

Step-3: Find the t-value

$$t = \frac{m}{\frac{s}{\sqrt{n}}} = \frac{58.125}{\frac{17.51}{\sqrt{8}}} = \frac{58.125}{6.191} = 9.38$$

Step-4: Comparison with the significance level

From step-3, we have
9.38 > 2.36
So, we have to reject the null hypothesis.
i.e. there is significant difference between BP reading before and after.

Conclusion:

T-test is a statistically significant test for the hypothesis testing (null and alternative hypotheses) when the sample size is small and the population parameter (mean and variance) is unknown.

3) Define F-Test?

An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled.

4) F-Test: (Explanation)

F-Test is any test that utilizes the F-Distribution table to fulfil its purpose (for eg: ANOVA). It compares the ratio of the variances of two populations and determines if they are statistically similar or not.

We can use this test when:

- The population is normally distributed.
- The samples are taken at random and are independent samples.

Formulas Used

where,

F_{calc} = Critical F-value.
σ_1^2 & σ_2^2 = variance of the two samples.

where,

df = Degrees of freedom of the sample.
n_s = Sample size.

Steps involved:
Step 1: Use Standard deviation (σ) and find variance (σ2) of the data. (if not already given)
Step 2: Determine the null and alternate hypothesis.

- H0 -> no difference in variances.
- Ha -> difference in variances.

Step 3: Find Fcalc using Eq-1.

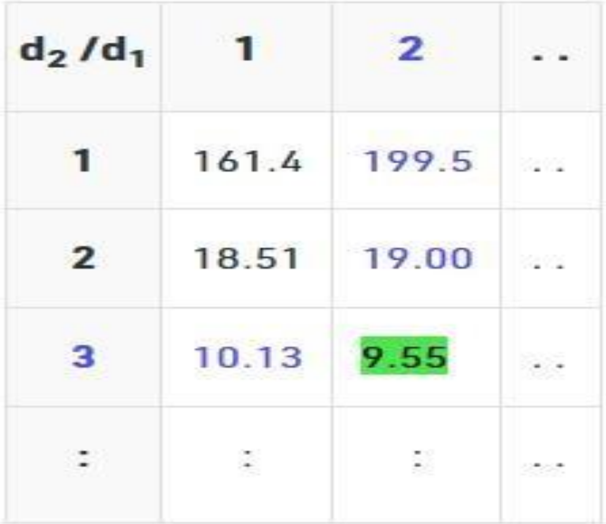
NOTE : *While calculating **F_{calc}**, divide the larger variance with small variance as it makes calculations easier.*
Step 4: Find the degrees of freedom of the two samples.
Step 5: Find Ftable value using d1 and d2 obtained in Step-4 from the F-distribution table. (link here). Take learning rate, α = 0.05 (if not given)
Looking up the F-distribution table:
In the F-Distribution table, refer the table as per the given value of α in the question.

- d1 (Across) = df of the sample with numerator variance. (larger)
- d2 (Below) = df of the sample with denominator variance. (smaller)

Consider the F-Distribution table given below,

While performing One-Tailed F-Test.
GIVEN :
α = 0.05

$d_1 = 2$
$d_2 = 3$



d_2 / d_1	1	2	..
1	161.4	199.5	..
2	18.51	19.00	..
3	10.13	9.55	..
:	:	:	...

Then, **F_{table} = 9.55**

Step 6: Interpret the results using *Fcalc* and *Ftable*.

Interpreting the results:

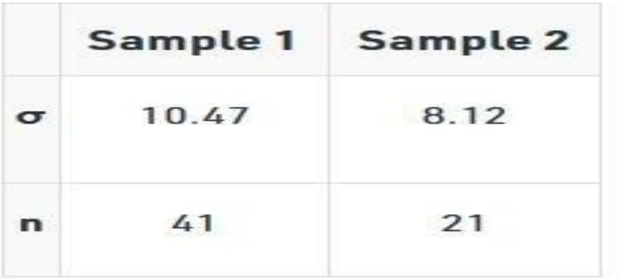
```
If Fcalc < Ftable :
    Cannot reject null hypothesis.
    ∴ Variance of two populations are similar.

If Fcalc > Ftable :
    Reject null hypothesis.
    ∴ Variance of two populations are not similar.
```

Example Problem (Step by Step)

Consider the following example,

Conduct a two-tailed F-Test on the following samples:



	Sample 1	Sample 2
σ	10.47	8.12
n	41	21

Step 1:

- **σ_1^2 =** $(10.47)^2 = 109.63$
- **σ_2^2 =** $(8.12)^2 = 65.99$

Step 2:

- **H_0:** no difference in variances.
- **H_a:** difference in variances.

Step 3:
F_{calc} = (109.63 / 65.99) = **1.66**

Step 4:
d_1 = (n_1 – 1) = (41 – 1) = 40
d_2 = (n_2 — 1) = (21 – 1) = 20

Step 5 - Using **d_1 = 40** and **d_2 = 20** in the F-Distribution table.
Take **α = 0.05** as it's not given.
Since it is a two-tailed F-test,
α = 0.05/2
= 0.025
Therefore, **F_{table} = 2.287**

Step 6 - Since $F_{calc} < F_{table}$ **(1.66 < 2.287)**:
We cannot reject null hypothesis.
∴ Variance of two populations are similar to each other.

F-Test is the most often used when comparing statistical models that have been fitted to a data set to identify the model that best fits the population. Researchers usually use it when they want to test whether two independent samples have been drawn from a normal population with the same variability.

5) What is analysis of variance?

Analysis of variance is a collection of statistical models and their associated estimation procedures used to analyze the differences among means. ANOVA was developed by the statistician Ronald Fisher

6) Define effect size estimation.

Effect size estimates provide important information about the impact of a treatment on the outcome of interest or on the association between variables. • Effect size estimates provide a common metric to compare the direction and strength of the relationship between variables across studies.

7) What is mean by multiple comparisons, multiplicity or multiple testing.

The multiple comparisons, multiplicity or multiple testing problem occurs when one considers a set of statistical inferences simultaneously or infers a subset of parameters selected based on the observed values. The more inferences are made, the more likely erroneous inferences become

8) Define ANOVA.

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

9) The Formula for ANOVA is:

$$\mathrm{F} = \frac{\mathrm{MST}}{\mathrm{MSE}}$$

where:

$\mathrm{F} = \text{ANOVA coefficient}$

$\mathrm{MST} = \text{Mean sum of squares due to treatment}$

$\mathrm{MSE} = \text{Mean sum of squares due to error}$

10) One-Way ANOVA vs. Two-Way ANOVA:

There are two main types of analysis of variance: one-way (or unidirectional) and two-way (bidirectional). One-way or two-way refers to the number of independent variables in your analysis of variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether the observed differences between the means of independent (unrelated) groups are explainable by chance alone, or whether there are any statistically significant differences between groups.

A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as department and gender. It is utilized to observe the interaction between the two factors. It tests the effect of two factors at the same time.

A three-way ANOVA, also known as three-factor ANOVA, is a statistical means of determining the effect of three factors on an outcome.

11) Mention a two-factor factorial design.

A two-factor factorial design is an experimental design in which data is collected for all possible combinations of the levels of the two factors of interest. If equal sample sizes are taken for each of the possible factor combinations then the design is a balanced two-factor factorial design.

12) Define statistical test in F-test.

An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled.

13) What are the two- way analyses of variance?

The two-way analysis of variance is an extension of the one-way ANOVA that examines the influence of two different categorical independent variables on one continuous dependent variable.

14) What are the types of ANOVA?

There are two main types of ANOVA: one-way (or unidirectional) and two-way. There also variations of ANOVA. For example, MANOVA (multivariate ANOVA) differs from ANOVA as the former tests for multiple dependent variables simultaneously while the latter assesses only one dependent variable at a time.

15) Define chi-square test.

The Chi-Square test is a statistical procedure used by researchers to examine the differences between categorical variables in the same population.

For example, imagine that a research group is interested in whether or not education level and marital status are related for all people in the U.S.

16) What Does the Analysis of Variance Reveal?

The ANOVA test is the initial step in analyzing factors that affect a given data set. Once the test is finished, an analyst performs additional testing on the methodical factors that measurably contribute to the data set's inconsistency. The analyst utilizes the ANOVA test results in an f- test to generate additional data that aligns with the proposed regression models.

The ANOVA test allows a comparison of more than two groups at the same time to determinewhether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.

If no real difference exists between the tested groups, which is called the null hypothesis, the result of the ANOVA's F-ratio statistic will be close to 1. The distribution of all possible values of the F statistic is the F-distribution. This is actually a group of distribution functions, with two characteristic numbers, called the numerator degrees of freedom and the denominator degrees offreedom.

17) How to Use ANOVA?

A researcher might, for example, test students from multiple colleges to see if students from one of the colleges consistently outperform students from the other colleges. In a business application, an R&D researcher might test two different processes of creating a product to see if one process is better than the other in terms of cost efficiency.

The type of ANOVA test used depends on a number of factors. It is applied when data needs to be experimental. Analysis of variance is employed if there is no access to statistical software resulting in computing ANOVA by hand. It is simple to use and best suited for small samples. With many experimental designs, the sample sizes have to be the same for the various factor level combinations.

ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources. It is employed with subjects, test groups, between groups and within groups.

18) What is the Analysis of Variance in Other Applications

In addition to its applications in the finance industry, ANOVA is also used in a wide variety of contexts and applications to test hypotheses in reviewing clinical trial data.For example, to compare the effects of different treatment protocols on patient outcomes; in social science research (for instance to assess the effects of gender and class on specified variables), in software engineering (for instance to evaluate database management systems), in manufacturing (to assess product and process quality metrics), and industrial design among other fields.

19) What is a Test?

In technical analysis and trading, a test is when a stock's price approaches an established support or resistance level set by the market. If the stock stays within the support and resistance levels, the test passes. However, if the stock price reaches new lows and/or new highs, the test fails. In other words, for technical analysis, price levels are tested to see if patterns or signals are accurate.

A test may also refer to one or more statistical techniques used to evaluate differences or similarities between estimated values from models or variables found in data. Examples includethe t-test and z-test

20) Define Range-Bound Market Test.

When a stock is range-bound, price frequently tests the trading range's upper and lower boundaries. If traders are using a strategy that buys support and sells resistance, they should wait for several tests of these boundaries to confirm price respects them before entering a trade.

Once in a position, traders should place a stop-loss order in case the next test of support or resistance fails.

21) What is the Trending Market Test?

In an up-trending market, previous resistance becomes support, while in a down-trending market, past support becomes resistance. Once price breaks out to a new high or low, it often retraces to test these levels before resuming in the direction of the trend. Momentumtraders can use the test of a previous swing high or swing low to enter a position at a morefavorable price than if they would have chased the initial breakout.

A stop-loss order should be placed directly below the test area to close the trade if the trend unexpectedly reverses.

22) Define Statistical Tests.

Inferential statistics uses the properties of data to test hypotheses and draw conclusions. Hypothesis testing allows one to test an idea using a data sample with regard to a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis. In particular, one seeks to reject the null hypothesis, or the notion that one or more random variables have no effect on another. If this can be rejected, the variables are likely to be associated with one another

23) What is Alpha Risk?

Alpha risk is the risk that in a statistical test a null hypothesis will be rejected when it is actuallytrue. This is also known as a type I error, or a false positive. The term "risk" refers to the chance or likelihood of making an incorrect decision. The primary determinant of the amount of alpha risk is the sample size used for the test. Specifically, the larger the sample tested, the lower the alpha risk becomes.

Alpha risk can be contrasted with beta risk, or the risk of committing a type II error (i.e., a falsenegative).

Alpha risk, in this context, is unrelated to the investment risk associated with an actively managed portfolio that seeks alpha, or excess returns above the market.

24) What is Range-Bound Trading?

Range-bound trading is a trading strategy that seeks to identify and capitalize on securities, like stocks, trading in price channels. After finding major support and resistance levels and connecting them with horizontal trendlines, a trader can buy a security at the lower trendline support (bottom of the channel) and sell it at the upper trendline resistance (top of the channel).

25) What is a One-Tailed Test?

A one-tailed test is a statistical test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both. If the sample being testedfalls into the one-sided critical area, the alternative hypothesis will be accepted instead of the null hypothesis.

AD3491 FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

UNIT V Notes

PART A

1. What Is Predictive Analytics?

The term predictive analytics refers to the use of statistics and modeling techniques to make predictions about future outcomes and performance. Predictive analytics looks at current and historical data patterns to determine if those patterns are likely to emerge again. This allows businesses and investors to adjust where they use their resources to take advantage of possible future events. Predictive analysis can also be used to improve operational efficiencies and reduce risk

2. Understanding Predictive Analytics.

Predictive analytics is a form of technology that makes predictions about certain unknowns in the future. It draws on a series of techniques to make these determinations, including artificial intelligence (AI), data mining, machine learning, modeling, and statistics.[3] For instance, data mining involves the analysis of large sets of data to detect patterns from it. Text analysis does the same, except for large blocks of text

3. Predictive models are used for all kinds of applications, including:

- Weather forecasts
- Creating video games
- Translating voice to text for mobile phone messaging
- Customer service
- Investment portfolio development

All of these applications use descriptive statistical models of existing data to make predictions about future data

4. What is mean by Forecasting?

Forecasting is essential in manufacturing because it ensures the optimal utilization of resources in a supply chain. Critical spokes of the supply chain wheel, whether it is inventory management or the shop floor, require accurate forecasts for functioning.

Predictive modelling is often used to clean and optimize the quality of data used for such forecasts. Modelling ensures that more data can be ingested by the system, including from customer-facing operations, to ensure a more accurate forecast.

5. Define Credit.

Credit scoring makes extensive use of predictive analytics. When a consumer or business applies for credit, data on the applicant's credit history and the credit record of borrowers with similar characteristics are used to predict the risk that the applicant might fail to perform on any credit extended.

6. Define Underwriting.

Data and predictive analytics play an important role in underwriting. Insurance companies examine policy applicants to determine the likelihood of having to pay out for a future claim based on the current risk pool of similar policyholders, as well as past events that have resulted in pay-outs. Predictive models that consider characteristics in comparison to data about past policyholders and claims are routinely used by actuaries.

7. What is mean by Marketing?

Individuals who work in this field look at how consumers have reacted to the overall economy when planning on a new campaign. They can use these shifts in demographics to determine if the current mix of products will entice consumers to make a purchase.

Active traders, meanwhile, look at a variety of metrics based on past events when deciding whether to buy or sell a security. Moving averages, bands, and breakpoints are based on historical data and are used to forecast future price movements

8. Predictive Analytics vs. Machine Learning

A common misconception is that predictive analytics and machine learning are the same things. Predictive analytics help us understand possible future occurrences by analyzing the past. At its core, predictive analytics includes a series of statistical techniques (including machine learning, predictive modelling, and data mining) and uses statistics (both historical and current) to estimate, or predict, future outcomes

9. What is the Decision Trees?

If you want to understand what leads to someone's decisions, then you may find decision trees useful. This type of model places data into different sections based on certain variables, such as price or market capitalization. Just as the name implies, it looks like a tree with individual branches and leaves. Branches indicate the choices available while individual leaves represent a particular decision.

Decision trees are the simplest models because they're easy to understand and dissect. They're also very useful when you need to make a decision in a short period of time

10. Define Regression.

This is the model that is used the most in statistical analysis. Use it when you want to determine patterns in large sets of data and when there's a linear relationship between the inputs. This method works by figuring out a formula, which represents the relationship between all the inputs found in the dataset. For example, you can use regression to figure out how price and other key factors can shape the performance of a security

11. Define Neural Networks.

Neural networks were developed as a form of predictive analytics by imitating the way the human brain works. This model can deal with complex data relationships using artificial intelligence and pattern recognition. Use it if you have several hurdles that you need to overcome like when you have too much data on hand, when you don't have the formula you need to help you find a relationship between the inputs and outputs in your dataset, or when you need to make predictions rather than come up with explanations.

12. What are the Benefits of Predictive Analytics?

There are numerous benefits to using predictive analysis. As mentioned above, using this type of analysis can help entities when you need to make predictions about outcomes when there are no other (and obvious) answers available.[9]

Investors, financial professionals, and business leaders are able to use models to help reduce risk. For instance, an investor and their advisor can use certain models to help craft an investment portfolio with minimal risk to the investor by taking certain factors into consideration, such as age, capital, and goals.[9]

There is a significant impact to cost reduction when models are used. Businesses can determine the likelihood of success or failure of a product before it launches. Or they can set aside capital for production improvements by using predictive techniques before the manufacturing process begins

13. Criticism of Predictive Analytics.

The use of predictive analytics has been criticized and, in some cases, legally restricted due to perceived inequities in its outcomes. Most commonly, this involves predictive models that result in statistical discrimination against racial or ethnic groups in areas such as credit scoring, home lending, employment, or risk of criminal behaviour.

A famous example of this is the (now illegal) practice of redlining in home lending by banks. Regardless of whether the predictions drawn from the use of such analytics are accurate, their use is generally frowned upon, and data that explicitly include information such as a person's race are now often excluded from predictive analytics.

14. How Does Netflix Use Predictive Analytics?

Data collection is very important to a company like Netflix. It collects data from its customers based on their behaviour and past viewing patterns. It uses information and makes predictions based to make recommendations based on their preferences. This is the basis behind the "Because you watched..." lists you'll find on your subscription.

15. What Is Data Analytics?

Data analytics is the science of analysing raw data to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption

16. What are the various steps of Data Analysis?

The process involved in data analysis involves several different steps:

1. The first step is to determine the data requirements or how the data is grouped. Data may be separated by age, demographic, income, or gender. Data values may be numerical or be divided by category.
2. The second step in data analytics is the process of collecting it. This can be done through a variety of sources such as computers, online sources, cameras, environmental sources, or through personnel.
3. Once the data is collected, it must be organized so it can be analyzed. This may take place on a spreadsheet or other form of software that can take statistical data.
4. The data is then cleaned up before analysis. This means it is scrubbed and checked to ensure there is no duplication or error, and that it is not incomplete. This step helps correct any errors before it goes on to a data analyst to be analyzed

PART B

1. How do you solve the least square problem in Python? What is least square method in Python?

2. What is the goodness-of-fit test?

Employers want to know which days of the week employees are absent in a five-day work week. Most employers would like to believe that employees are absent equally during the week. Suppose a random sample of 60 managers were asked on which day of the week they had the highest number of employee absences. The results were distributed as in the table below. For the population of employees, do the days for the highest number of absences occur with equal frequencies during a five-day work week? Test at a 5% significance level.

Day of the Week Employees were Most Absent

	Monday	**Tuesday**	**Wednesday**
Number of Absences	15	12	9

3. One study indicates that the number of televisions that American families have is distributed (this is the **given** distribution for the American population) as in the table.

Number of Televisions	**Percent**
	10
1	16
2	55
3	11
4+	8

The table contains expected (*E*) percents.

A random sample of 600 families in the far western United States resulted in the data in this table.

Number of Televisions	**Frequency**
0	66

Number of Televisions	**Frequency**
1	119
2	340
3	60
4+	15 ***Total = 600***

The table contains observed (*O*) frequency values.

At the 1% significance level, does it appear that the distribution “number of televisions” of far western United States families is different from the distribution for the American population as a whole?

4. Explain in detail about time series analysis with example.

5. Describe Regression using Stats Models.

6. Explain multiple regression with an example.

7. What is the nonlinear relationships and types .Difference between linear and non linear relationship

8. Describe logistic regression in detail.

9. Explain in detail serial correlation and autocorrelation.

https://www.investopedia.com/terms/s/serial-correlation.asp#:~:text=Serial%20correlation%20is%20the%20relationship,it%20may%20not%20be%20random.

https://corporatefinanceinstitute.com/resources/data-science/autocorrelation/

10) Describe in detail Introduction to survival analysis.