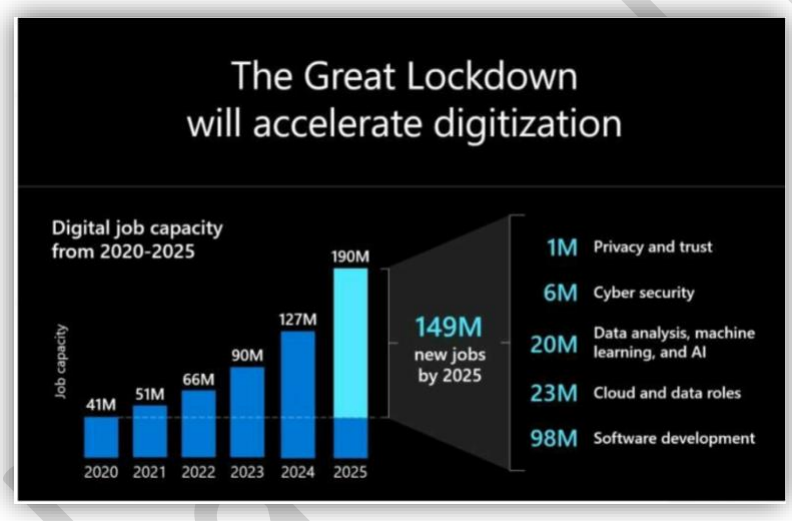
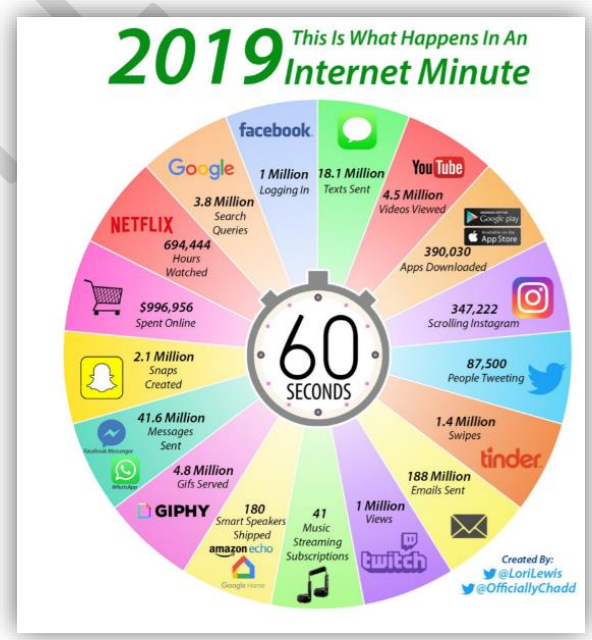
Business Analytics

Chapter objectives:

- Define business analytics
- Explain the relationship of analytics and business intelligence to the subject of business analytics
- Describe the three steps of the business analytics process
- Describe four data classification measurement scales
- Explain the relationship of the business analytics process with the organization decision-making process





Introduction to Business Analytics

Business analytics begins with a data set. It is a simple collection of data or a data file.

'Database' (a collection of data files that contain information on people, locations, and so on). As databases grow, they need to be stored somewhere. Technologies such as computer clouds (hardware and software used for data remote storage, retrieval, and computational functions)

'Data Warehousing' (a collection of databases used for reporting and data analysis) store data. Database storage areas have become so large that a new term was devised to describe them.

'Big Data' describes the collection of data sets that are so large and complex that software systems are hardly able to process them (7 V's)

1. Volume
2. Variety
3. Velocity
4. Veracity
5. Variability
6. Visualization
7. Value

Three terms in business literature are often related to one another: **analytics, business analytics, and business intelligence.**

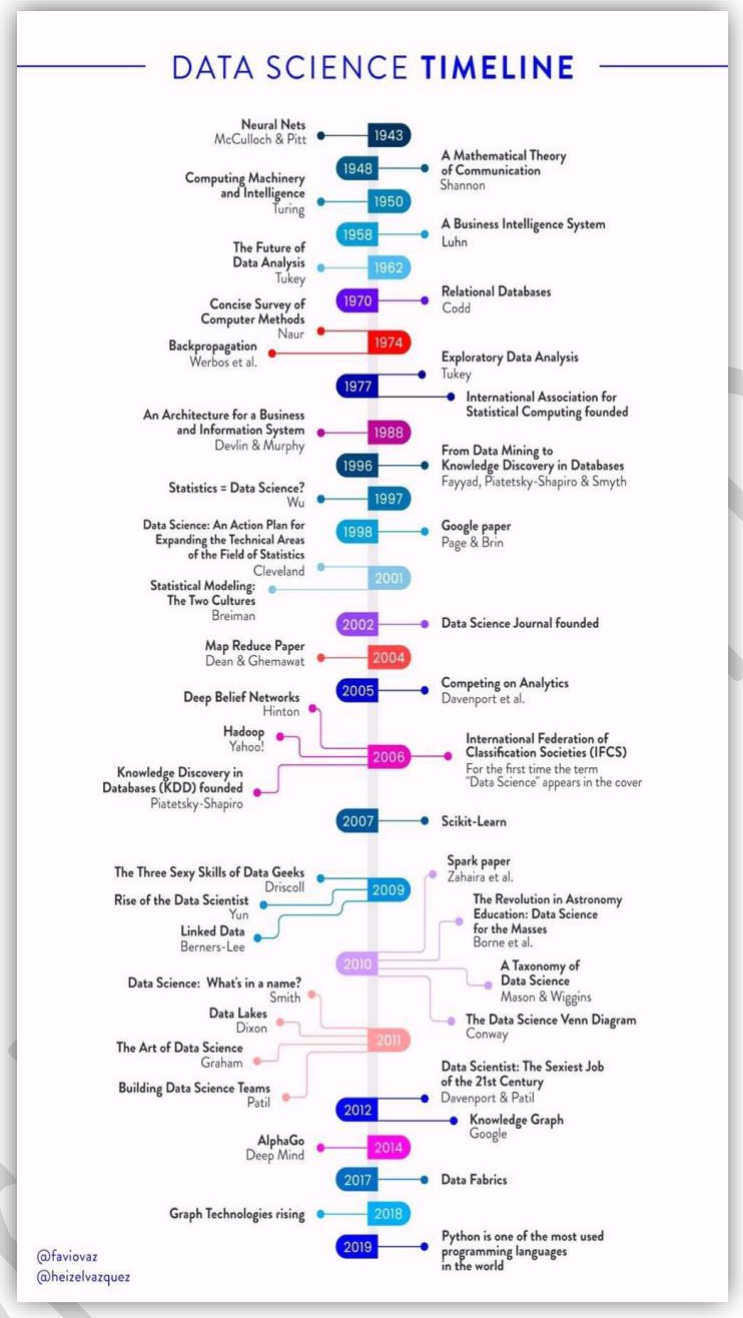
1. **Domain/Business Intelligence (In-Out)**
2. **Business Scenario**
3. **Problem Statement**
4. **Data (Data is ready to speak with you, Are you ready listen)**
5. **Model/Algorithm**
6. **Statistics, Probability, Linear Algebra, Calculus**
7. **Language (Python, R, Julia, SAS, Java, Scala, Matlab)**
8. **Model Building, Validation Matrices, Model improvement, Re validate**
9. **Change the model**

10. **Results and Interpretation**
11. **Action**

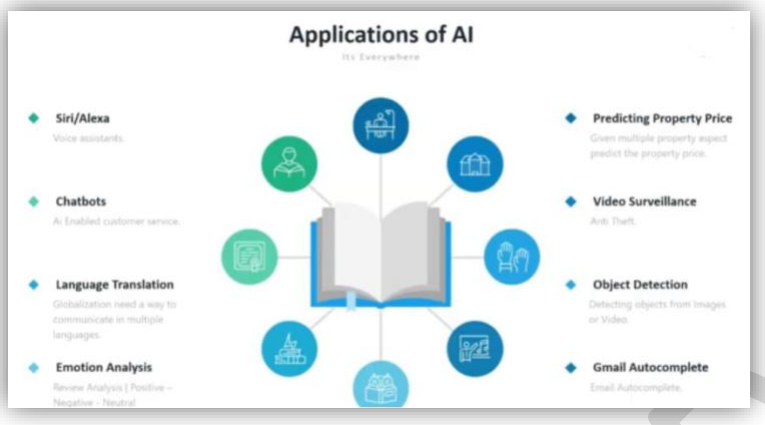
Analytics can be defined as a process that involves the use of **statistical techniques** (measures of central tendency, graphs, and so on), **information system software** (data mining, sorting routines), and **operations research methodologies** (linear programming) to explore, visualize, discover, and communicate patterns or trends in data.

Simply, **analytics converts data into useful information.**

Analytics is an older term commonly applied to all disciplines, not just business. A typical example of the use of analytics is the weather measurements collected and converted into statistics, which in turn predict weather patterns.


DATA SCIENCE TIMELINE
Neural Nets
McCulloch & Pitt
1943
1948
A Mathematical Theory of Communication
Shannon
Computing Machinery and Intelligence
Turing
1950
1958
A Business Intelligence System
Luhn
The Future of Data Analysis
Tukey
1962
1970
Relational Databases
Codd
Concise Survey of Computer Methods
Naur
1974
Backpropagation
Werbos et al.
Exploratory Data Analysis
Tukey
1977
International Association for Statistical Computing founded
An Architecture for a Business and Information System
Devlin & Murphy
1988
1996
From Data Mining to Knowledge Discovery in Databases
Fayyad, Piatetsky-Shapiro & Smyth
Statistics = Data Science?
Wu
1997
Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics
Cleveland
1998
Google paper
Page & Brin
2001
Statistical Modeling: The Two Cultures
Breiman
2002
Data Science Journal founded
Map Reduce Paper
Dean & Ghemawat
2004
2005
Competing on Analytics
Davenport et al.
Deep Belief Networks
Hinton
Hadoop
Yahoo!
2006
International Federation of Classification Societies (IFCS)
For the first time the term "Data Science" appears in the cover
Knowledge Discovery in Databases (KDD) founded
Piatetsky-Shapiro
2007
Scikit-Learn
Spark paper
Zahaira et al.
The Three Sexy Skills of Data Geeks
Driscoll
2009
Rise of the Data Scientist
Yun
The Revolution in Astronomy Education: Data Science for the Masses
Borne et al.
Linked Data
Berners-Lee
2010
A Taxonomy of Data Science
Mason & Wiggins
Data Science: What's in a name?
Smith
Data Lakes
Dixon
The Data Science Venn Diagram
Conway
2011
The Art of Data Science
Graham
Data Scientist: The Sexiest Job of the 21st Century
Davenport & Patil
Building Data Science Teams
Patil
2012
Knowledge Graph
Google
AlphaGo
Deep Mind
2014
2017
Data Fabrics
Graph Technologies rising
2018
2019
Python is one of the most used programming languages in the world
@faviovaz
@heizelvazquez


Text – Image – Audio - Video
Text
Language Translation
Text Summarization
Chatbots
Sentiment Analysis
Search Engines
Document Similarity
Image
Image Search
Object Detection
Image Similarity
Image Captioning
Image Generation
Image Coloring
Audio
Speech Analysis
Audio Recognition
Alexa / Siri
Voice Control
Language Translation
Noise Cancellation
Video
Video Summarization
Object Detection
Activity Recognition
Emotion Analysis
Video Captioning
Video Search



There are many types of analytics, and there is a need to organize these types to understand their uses. We will adopt the three categories (descriptive, predictive, and prescriptive).

These types of analytics can be viewed independently.

The Institute of Operations Research and Management Sciences (INFORMS) organization (www.informs.org) suggests for grouping the types of analytics (see T able 1 .1)

Table 1.1 Types of Analytics

Type of Analytics	Definition
Descriptive	The application of simple statistical techniques that describe what is contained in a data set or database. Example: An age bar chart is used to depict retail shoppers for a department store that wants to target advertising to customers by age.
Predictive	An application of advanced statistical, information software, or opera research methods to identify predictive variables and build predictiv models to identify trends and relationships not readily observed in a descriptive analysis. Example: Multiple regression is used to show th relationship (or lack of relationship) between age, weight, and exerci diet food sales. Knowing that relationships exist helps explain why or of independent variables influences dependent variables such as bus performance.
Prescriptive	An application of decision science, management science, and operations research methodologies (applied mathematical techniques) to make best use of allocable resources. Example: A department store has a limited advertising budget to target customers. Linear programming models can be used to optimally allocate the budget to various advertising media.

Linear Regression (Quantitative)

Logistic Regression (Qualitative)

The process of analytics can involve any one of the three types of analytics; the major components of business analytics include all three used in combination to generate new, unique, and valuable information that can aid business organization decision-making.

In addition, the three types of analytics are applied sequentially (descriptive, then predictive, then prescriptive). Therefore, business analytics (BA) can be defined as a process beginning with business-related data collection and consisting of sequential application of descriptive, predictive, and prescriptive major analytic components, the outcome of which supports and demonstrates business decision-making and organizational performance.

Finally, 'Business Analytics' goes beyond plain analytics, requiring a clear relevancy to business, a resulting insight that will be implementable, and performance and value measurement to ensure a successful business result.

The purposes and methodologies used for each of the three types of analytics differ, as can be seen in T able 1 .2.

Table 1.2 Analytic Purposes and Tools

Type of Analytics	Purpose	Examples of Methodologies
Descriptive	To identify possible trends in large data sets or databases. The purpose is to get a rough picture of what generally the data looks like and what criteria might have potential for identifying trends or future business behavior.	Descriptive statistics, including measures of central tendency (mean, median, mode), measures of dispersion (standard deviation), charts, graphs, sorting methods, frequency distributions, probability distributions, and sampling methods.
Predictive	To build predictive models designed to identify and predict future trends.	Statistical methods like multiple regression and ANOVA. Information system methods like data mining and sorting. Operations research methods like forecasting models.
Prescriptive	To allocate resources optimally to take advantage of predicted trends or future opportunities.	Operations research methodologies like linear programming and decision theory.

Analytics, BA, and BI will be mentioned throughout this book. A review of characteristics to help differentiate these terms is presented in Table 1.3 .

Table 1.3 Characteristics of Analytics, Business Analytics, and Business Intelligence

Characteristics	Analytics	Business Analytics (BA)	Business Intelligence (BI)
Business performance planning role	What is happening, and what will be happening?	What is happening now, what will be happening, and what is the best strategy to deal with it?	What is happening now, and what have we done in the past to deal with it?
Use of descriptive analytics as a major component of analysis	Yes	Yes	Yes
Use of predictive analytics as a major component of analysis	Yes	Yes	No (only histor
Use of prescriptive analytics as a major component of analysis	Yes	Yes	No (only historically)
Use of all three in combination	No	Yes	No
Business focus	Maybe	Yes	Yes
Focus of storing and maintaining data	No	No	Yes
Required focus of improving business value and performance	No	Yes	No

1. Data Preprocessing
2. Data Cleaning
3. Exploratory Analysis (60-80%)

Business Analytics Process

The complete business analytics process involves the three major component steps applied sequentially to a source of data (see Figure 1.1). The outcome of the business analytics process must relate to business and seek to improve business performance in some way.

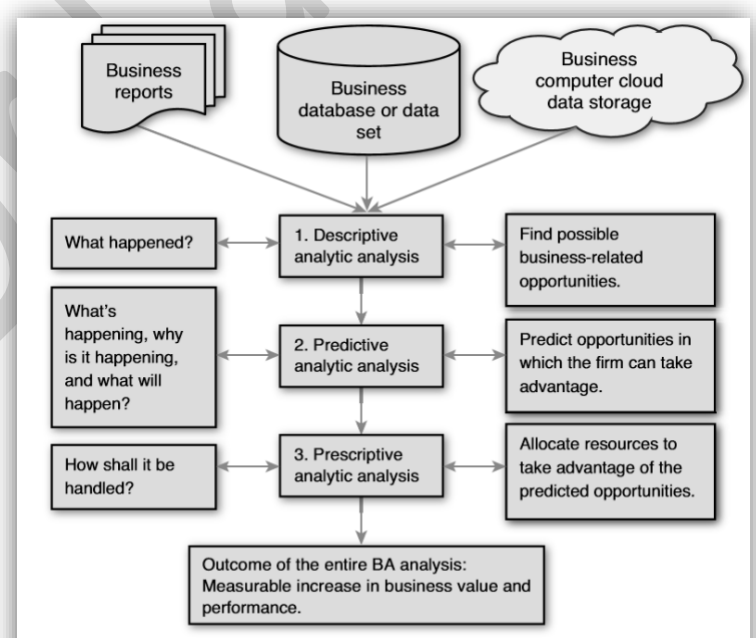


Figure 1.1 Business analytics process

Types of Data Measurement & Classification Scales

Numerical – Integer, Float

Categorical – String

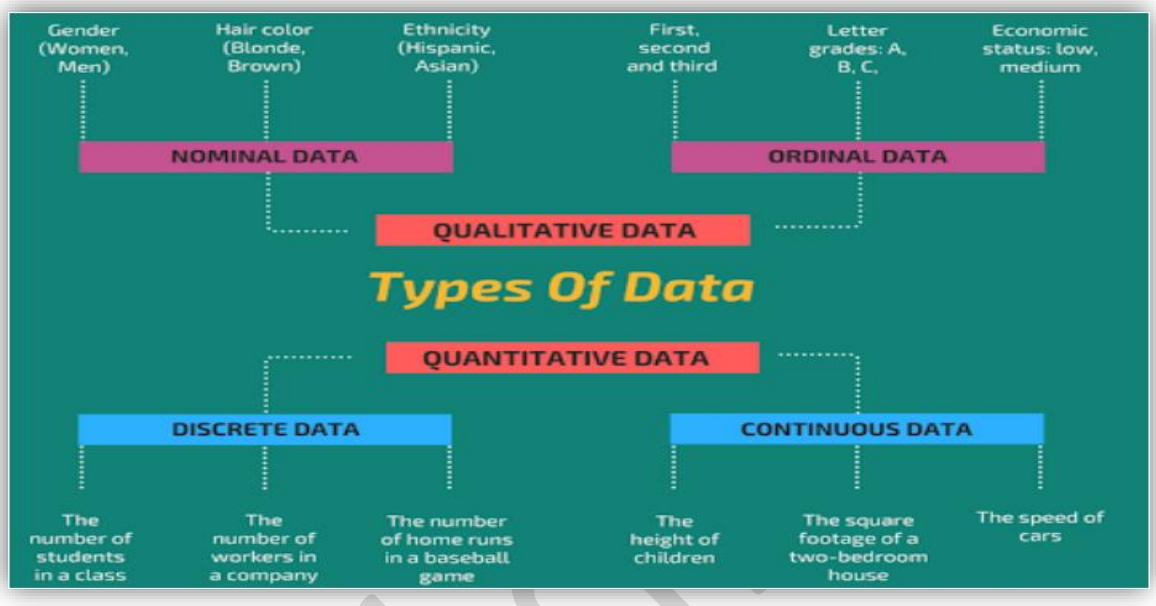


Table 1.4 Types of Data Measurement Classification Scales

Type of Data Measurement Scale	Description
Categorical Data	Data that is grouped by one or more characteristics. Categorical data usually involves cardinal numbers counted or expressed as percentages. Example 1: Product markets that can be characterized by categories of "high-end" products or "low-income" products, based on dollar sales. It is common to use this term to apply to data sets that contain items identified by categories as well as observations summarized in cross-tabulations or contingency tables.
Ordinal Data	Data that is ranked or ordered to show relational preference. Example 1: Football team rankings not based on points scored but on wins. Example 2: Ranking of business firms based on product quality.
Interval Data	Data that is arranged along a scale, in which each value is equally distant from others. It is ordinal data. Example 1: A temperature gauge. Example 2: A survey instrument using a Likert scale (that is, 1, 2, 3, 4, 5, 6, 7), where 1 to 2 is perceived as equidistant to the interval from 2 to 3, and so on. Note: In ordinal data, the ranking of firms might vary greatly from first place to second, but in interval data, they would have to be relationally proportional.
Ratio Data	Data expressed as a ratio on a continuous scale. Example 1: The ratio of firms with green manufacturing programs is twice that of firms without such a program.

Relationship of BA Process and Organization Decision-Making Process

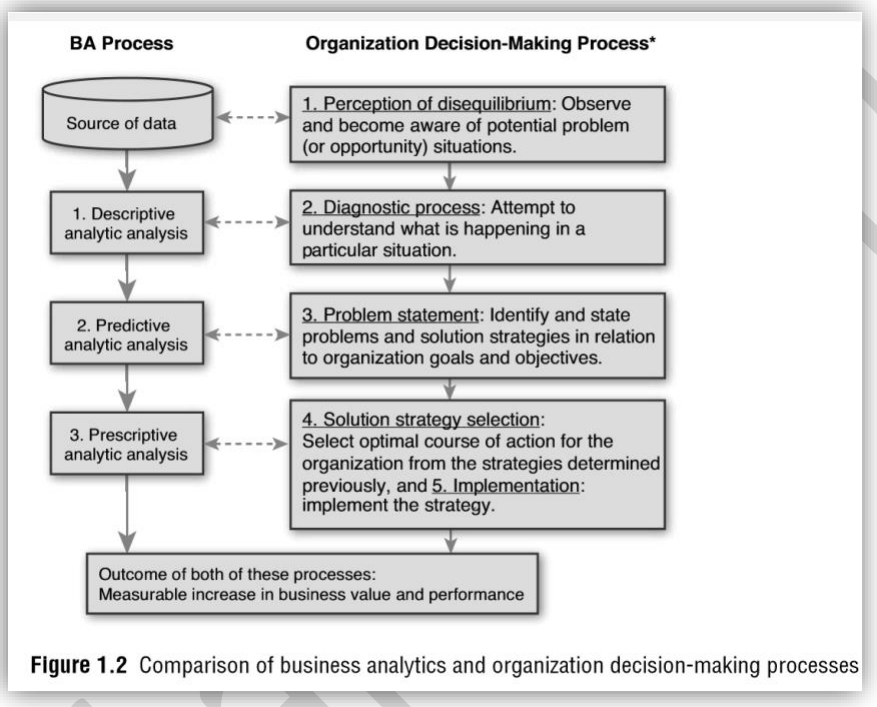
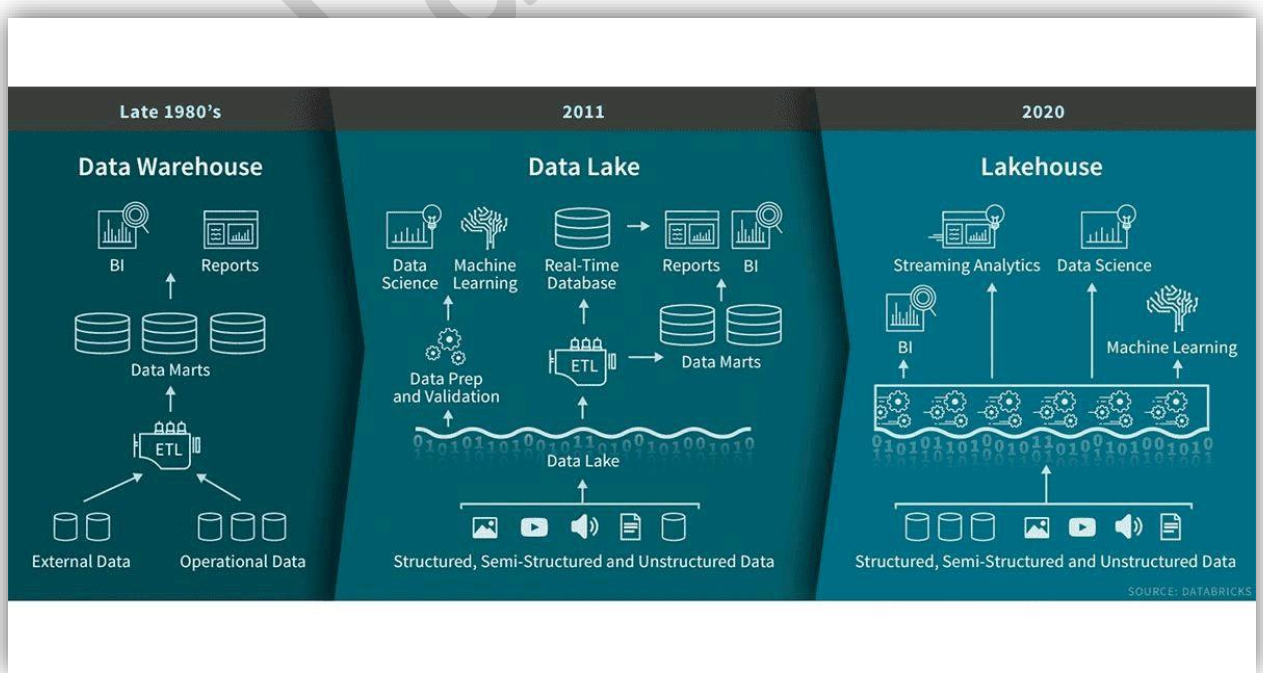


Figure 1.2 Comparison of business analytics and organization decision-making processes



'Next best action' is used in predicting the future business profitability.

The organization decision-making process (ODMP)

The five-step ODMP begins with the perception of disequilibrium, or the awareness that a problem exists that needs a decision.

Similarly, in the BA process, the first step is to recognize that databases may contain information that could both solve problems and find opportunities to improve business performance.

Then in Step 2 of the ODMP, an exploration of the problem to determine its size, impact, and other factors is undertaken to diagnose what the problem is.

Likewise, the BA descriptive analytic analysis explores factors that might prove useful in solving problems and offering opportunities.

The ODMP problem statement step is similarly structured to the BA predictive analysis to find strategies, paths, or trends that clearly define a problem or opportunity for an organization to solve problems.

Finally, the ODMP's last steps of strategy selection and implementation involve the same kinds of tasks that the BA process requires in the final prescriptive step.

Discussion Questions

1. What is the difference between analytics and business analytics?
2. What is the difference between business analytics and business intelligence?
3. Why are the steps in the business analytics process sequential?
4. How is the business analytics process similar to the organization decision making process?
5. Why does interval data have to be relationally proportional?

Competitive Advantages of Business Analytics

Find more sales opportunities to create competitive advantage in business

- Sales managers are able to quickly identify which customers are buying and, more importantly,
- What products are in decline?
- With BI, you can quickly visualize customer spending trends by monitoring their purchases on a daily, weekly, monthly, or annual basis.

Identifying these trends highlights new sales opportunities.

An important opportunity lies with cross-selling complementary products.

Decrease the number of customers you lose and create competitive advantage in business

You can monitor how much your customers have purchased on a daily, weekly, monthly or annual basis to identify their spending trends.

Providing your sales team with access to customer data such as buying patterns, previous feedback and behavior, offers clear insights into what your customers truly want and the ability to quickly spot customers on the decline. Armed with this information, your sales team can reach out before it's too late.

Most customers want to be heard. By having meaningful conversations with them, you demonstrate that you are listening. Feeling heard leads to customer satisfaction and retention. With access to Phocas from mobile devices, your reps can quickly respond to customers' needs even when they are out of the office.

Web result with site links

Phocas Software | Business intelligence and data analytics

www.phocassoftware.com

Pocas offers Data analytics and BI strategy for manufacturers, distributors and retailers.

Visualizations, KPIs and metrics to make your data accessible and useful

Measure over and under-stock and create competitive advantage in business

- A significant way to reduce wasteful spending is to streamline your inventory management system. The goal is to have enough stock on hand to fulfill customer orders, without overstocking.
- Overstocking ties up capital that can be best used elsewhere.
- Over time it is easy to see how much supply you need on hand.

Avoid dead stock and create competitive advantage in business

- Another danger of over-stock is the risk of it becoming dead stock.
- Dead stock is stock that remains on the shelf for so long that it cannot be sold or returned. Dead stock can be caused by a product losing popularity, or by the development of newer versions of the product making your stock is outdated.
- Dead stock can also be stock that has literally perished. With dead stock, not only has the investment capital been lost, but the longer it sits in your warehouse is wasted space for profitable products.

Gain a clearer view of the profitability of your rebates program

- Many rebate programs are complex, making them difficult to track. Many businesses use spreadsheets which are known for the potential for costly errors. When margins are tight, the ability to account for every dollar is critical.
- Now, you have a clear, easy-to-understand view of your profitability, enabling you to make better decisions.

Decrease the time IT is tied up with reporting requests and create competitive advantage in business

- In traditional settings, the IT department generates reports on request. These reports enable managers to see what's happening in their domains. However, these reports are static snapshots that cannot provide insight into the underlying factors driving the information. If more insight is needed, a manager must request another report.
- IT generated reports are also time consuming. The average turnaround time is a few days and up to a week. In today's fast past world, many factors may have changed by the time the report is returned, making it obsolete.

Increase the speed and accuracy of reporting and create competitive advantage in business

With BI, reporting is not only faster because reports don't need to go through IT, but also because it eliminates the need for error-prone spreadsheets.

Discover trends you did not know existed and create competitive advantage in business

Most executives have a pretty sound idea of the overall shape of their business. However, many find that after they implement a BI solution they discover new information. Having all of the information allows you to take advantage of unidentified opportunities and to address unrecognized problems before they can have a serious impact.

Discover new ways to cut costs

Organizations use cost-cutting to improve their competitive advantage by lowering production costs and passing on the benefits to consumers.

By merging data analytics and competitive advantage-focused strategies, businesses can leverage their monumental data assets given that they provide organizations with new ways of cost-cutting.

Analytics can reveal inefficiencies and flaws that would be impossible to capture through conventional means. Thus, organizations can develop new cost-cutting measures focusing on removing inefficiencies and cutting down waste in resources.

An excellent example is Airbus. The aviation corporation used data analytics to improve manufacturing, development and maintenance processes. The change in processes reduced the number of operational interruptions disrupting flight schedules and made better use of resources when designing a new aircraft. Data analytics also provided real-time data which was also expected to cut fuel usage by 15%.

Personalized customer service

The internet has changed the customer-retailer relationship. Thanks to Amazon and Netflix, customers have grown accustomed to personalized service and now expect organizations to provide a similar level of customer service. Those who don't will lose competitive advantage.

Fortunately, data analytics allows organizations to provide personalized customer service. Analytics platforms can capture and breakdown customer purchase history into a detailed individual profile. Analytics goes beyond the basic demographic information – it can even go as far as to depict what the customer will buy, and under what circumstances. This paves the way for personalized customer campaigns in brand awareness and engagement.

Amazon and Netflix are the best examples of organizations using data and analytics to provide personalized customer service to their customers. Using past purchases as a barometer, Amazon provides a series of recommendations that are catered specifically to that customer. Something that can only be achieved with data analytics because they have a shopper base consisting of millions.

UBER/OLA

1. **Car Pooling (0.5 Lacs)**

Pin location
Historical (past 3 weeks, past 3 months, 1 years)
Streaming Analytics

Search – mini, macro, prime, pooling, auto (closer)
Contact (Passenger – Driver) – Passenger Denial - Recording

Dynamic price - Historical (past 3 weeks, past 3 months, 1 years) – Demand – Competitors – Forecasting Analytics

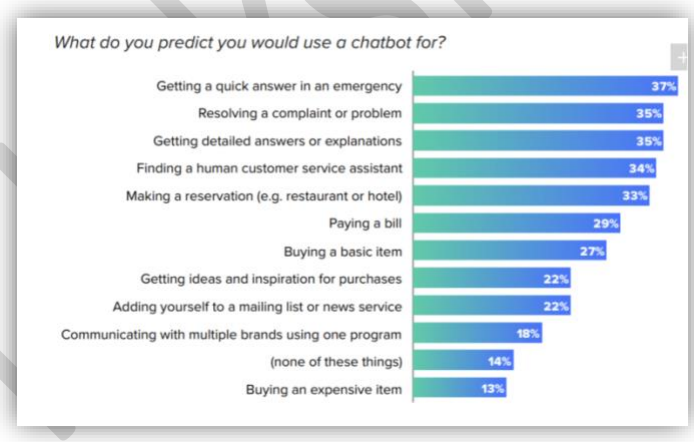
OTP

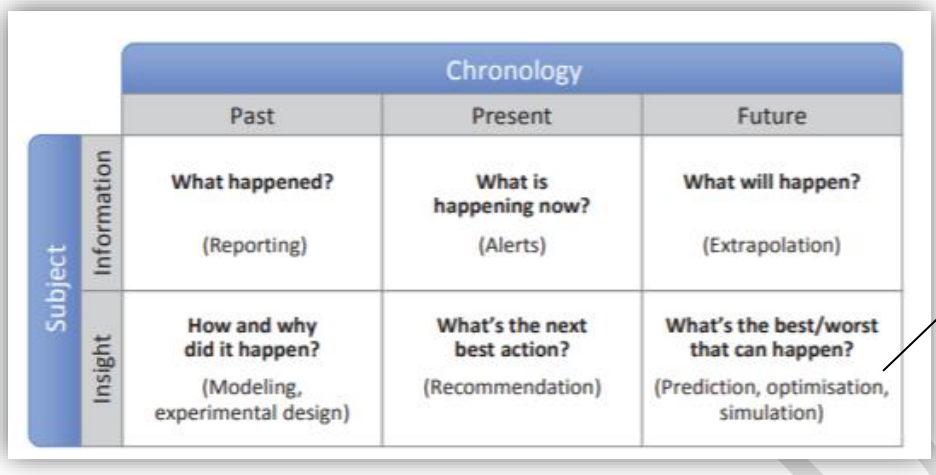
2. **Shortest Path (Map processing) – Traffic**

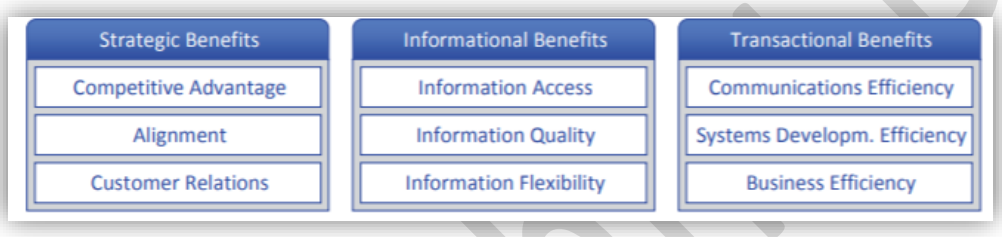
Signal tracking

Drop – rate, details,

Chatbot







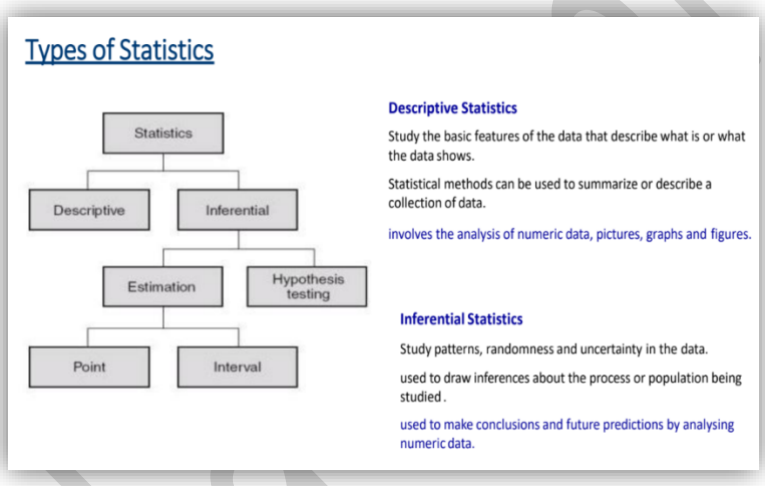
Benefits

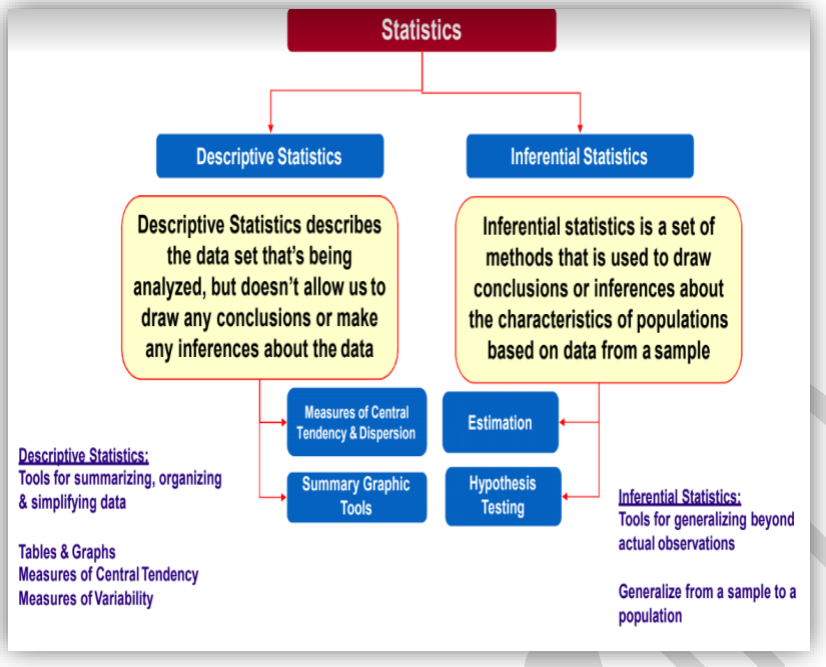
- Faster reporting, analysis or planning
- More accurate reporting, analysis or planning
- Better business decisions
- Improved data quality
- Improved employee satisfaction
- Improved operational efficiency
- Improved customer satisfaction
- Increased competitive advantage
- Reduced costs
- Increased revenues
- Saved headcount

Key takeaways

Competitive advantage is vital for any organization to survive. To compete, organizations have to discover new ways to cut costs, deploy resources more effectively, and develop ways to reach out to each of their customers individually. All of this can be achieved with data analytics. Data analytics provides brand new insights not found through other tools. These are insights that can pave the way for other organizations to make better use of their resources, cut costs and personalize their services to strengthen their advantage.

Descriptive Statistical methods





Descriptive statistics allow you to characterize your data based on its properties. There are four major types of descriptive statistics:

1. Measures of Frequency:

* Count, Percent, Frequency
* Shows how often something occurs
* Use this when you want to show how often a response is given

2. Measures of Central Tendency

* **Mean, Median, and Mode**
* Locates the distribution by various *points*
* Use this when you want to show how an average or most commonly indicated response

3. Measures of Dispersion or Variation

* **Range, Variance, Standard Deviation**
* Identifies the spread of scores by stating intervals
* Range = High/Low points
* Variance or Standard Deviation = difference between observed score and mean
* Use this when you want to show how "spread out" the data are. It is helpful to know when your data are so spread out that it affects the mean

DA

A

65
69
90
87
56

B

34
40
12
99
0

4. Measures of Position

* Percentile Ranks, Quartile Ranks
* Describes how scores fall in relation to one another. Relies on standardized scores
* Use this when you need to compare scores to a normalized score (e.g., a national norm)

Estimators (Cont.)

- Point estimate-
  - The mean annual rainfall of Melbourne is 620mm per year
- Interval Estimate-
  - In 80% of all years Melbourne receives between 440 and 800 mm rain

Descriptive Statistics

Descriptive Statistics: Methods of organizing, summarizing, and presenting data in an informative way.

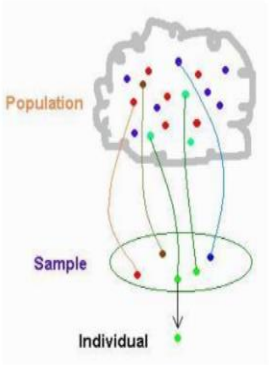
EX 1: A Gallup poll found that 49% of the people in a survey knew the name of the first book of the Bible. The statistic 49 describes the number out of every 100 persons who knew the answer.

EX 2: According to Consumer Reports, General Electric washing machine owners reported 9 problems per 100 machines during 2001. The statistic 9 describes the number of problems out of every 100 machines.

Inferential Statistics

Inferential Statistics: A decision, estimate, prediction, or generalization about a population, based on a sample.

A **Population** is a **Collection** of all possible individuals, objects, or measurements of interest.



A **Sample** is a portion, or part, of the population of interest

Examples of inferential statistics

Example 1: TV networks constantly monitor the popularity of their programs by hiring Nielsen and other organizations to sample the preferences of TV viewers.

Example 2: Wine tasters sip a few drops of wine to make a decision with respect to all the wine waiting to be released for sale.

Example 3: The accounting department of a large firm will select a sample of the invoices to check for accuracy for all the invoices of the company.

Are these sufficient?

- There is the man who drowned crossing a stream with an average depth of six inches. ~W.I.E. Gates
- Say you were standing with one foot in the oven and one foot in an ice bucket. According to the averages, you should be perfectly comfortable.

e.g. x1, x2, x3 Are the times taken to get to Delhi in different modes of transport

	Auto	Office Transport	Own Car
	7	9	1
	6	9	3
	3	9	5
	8	9	7
	12	9	9
	9	9	9
	9	9	9
	13	9	11
	13	9	13
	9	9	15
	10	9	17
Mean	9	9	9
Median	9	9	9
Mode	9	9	9

NO!!!

Now try this...

	Auto	Office Transport	Own Car
	7	9	1
	6	9	3
	3	9	5
	8	9	7
	12	9	9
	9	9	9
	9	9	9
	13	9	11
	13	9	13
	9	9	15
	10	9	17
Mean	9	9	9
Median	9	9	9
Mode	9	9	9

Std Dev	**3.0**	**0.0**	**4.9**
Variance	**9.2**	**0.0**	**24.0**

Measures of Dispersion (Variance)

Dispersion refers to the spread or variability in the data.

It determines how spread out are the scores around the mean.

The basic question being asked is how much do the scores deviate around the Mean? The more "bunched up" around the mean the better your ability to make accurate predictions.

Coefficient of Variation

The **coefficient of variation (CV)** is a normalized measure of dispersion of a probability distribution. It is defined as the ratio of the standard deviation to the mean :

- Measure of *relative* dispersion
- Always a %
- Shows variation relative to mean
- Used to compare 2 or more groups

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

Which Cricketer do you like? Who is more consistent?

Dravid	150	150	130	125	145	110	100	152	120	50	128
Sehwag	230	240	150	50	173	23	20	300	45	1	128

	Dravid	Sehwag
Mean	123.636	123.636
Median	128	128
CV	24%	84%

Inter quartile Range

- The *inter quartile range* is an **alternative measure of dispersion** that is less influenced than the range by **extreme values.**
- If (*nk*)/100 is an integer (a round number with no decimal or fractional part), the *k*th percentile of the observations is the average of the ((*nk*)/100)th and ((*nk*)/100 + 1)th largest observations.
- (*n* = the number of observations, *k* the percentile you wish to find)

Review of probability distribution and data modeling

Random Variable (Discrete (Normal) /Continuous R.V. follows probability distribution)

Random variable - The event that has more than one value (each value is associated with its probability/associated likelihood)

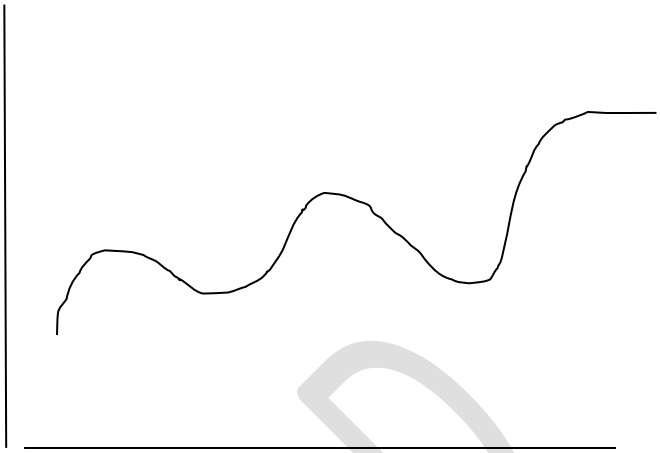
0.5 0

0.2 1 P(X=1) =0.2

0.3 2

Types of Distributions

1. Bernoulli Distribution
2. Uniform Distribution
3. Binomial Distribution
4. Normal Distribution
5. Poisson Distribution
6. Exponential Distribution



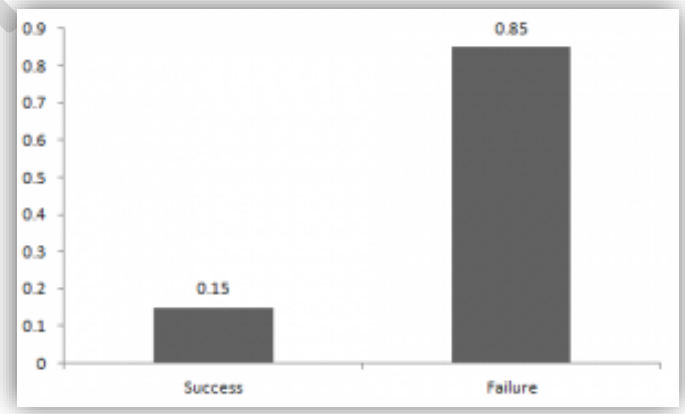
A **Bernoulli distribution** has only two possible outcomes, namely 1 (success) and 0 (failure), and a single trial. So the random variable X which has a Bernoulli distribution can take value 1 with the probability of success, say p, and the value 0 with the probability of failure, say q or 1-p.

Here, the occurrence of a head denotes success, and the occurrence of a tail denotes failure. Probability of getting a head = 0.5 = Probability of getting a tail since there are only two possible outcomes.

The probability mass function is

$$P(x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \end{cases}$$

The probabilities of success and failure need not be equally likely, like the result of a game between team 1 and team 2. Team 2 is pretty much certain to win. So in this case probability of team 1 success is 0.15 and their probability of failure is 0.85



The expected value of a random variable X from a Bernoulli distribution is found as follows:

$E(X) = 1*p + 0*(1-p) = p$

The variance of a random variable from a bernoulli distribution is:

$V(X) = E(X^2) – [E(X)]^2 = p – p^2 = p(1-p)$

There are many examples of Bernoulli distribution such as whether it's going to rain tomorrow or not where rain denotes success and no rain denotes failure and Winning (success) or losing (failure) the game.

Uniform Distribution

When you roll a fair die, the outcomes are 1 to 6. The probabilities of getting these outcomes are equally likely and that is the basis of a uniform distribution. Unlike Bernoulli Distribution, all the n number of possible outcomes of a uniform distribution are equally likely.

A variable X is said to be uniformly distributed if the density function is:

$$f(x) = \frac{1}{b-a} \qquad for\ -\infty < a \le x \le b < \infty$$

The graph of a uniform distribution curve looks like

Bus Arrival 9.00AM to 9.30AM

You can see that the shape of the Uniform distribution curve is rectangular, the reason why Uniform distribution is called rectangular distribution.

For a Uniform Distribution, a and b are the parameters.

The number of bouquets sold daily at a flower shop is uniformly distributed with a maximum of 40 and a minimum of 10.

Let's try calculating the probability that the daily sales will fall between 15 and 30.

The probability that daily sales will fall between 15 and 30 is (30-15)*(1/(40-10)) = 0.5

Similarly, the probability that daily sales are greater than 20 is = 0.667

The mean and variance of X following a uniform distribution is:

Mean -> E(X) = (a+b)/2

Variance -> V(X) = $(b-a)^2/12$

The standard uniform density has parameters a = 0 and b = 1, so the PDF for standard uniform density is given by:

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & otherwise \end{cases}$$

Binomial Distribution

Suppose that you won the toss today and this indicates a successful event. You toss again but you lost this time. If you win a toss today, this does not necessitate that you will win the toss tomorrow. Let's assign a random variable, say X, to the number of times you won the toss. What can be the possible value of X? It can be any number depending on the number of times you tossed a coin.

There are only two possible outcomes. Head denoting success and tail denoting failure. Therefore, probability of getting a **head = 0.5 and the probability of failure can be easily computed as: q = 1- p = 0.5.**

A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials is called a Binomial Distribution.

Each trial is independent since the outcome of the previous toss doesn't determine or affect the outcome of the current toss. An experiment with only two possible outcomes repeated n number of times is called binomial. The parameters of a binomial distribution are n and p where n is the total number of trials and p is the probability of success in each trial.

On the basis of the above explanation, the properties of a Binomial Distribution are

Each trial is independent.

There are only two possible outcomes in a trial- either a success or a failure.

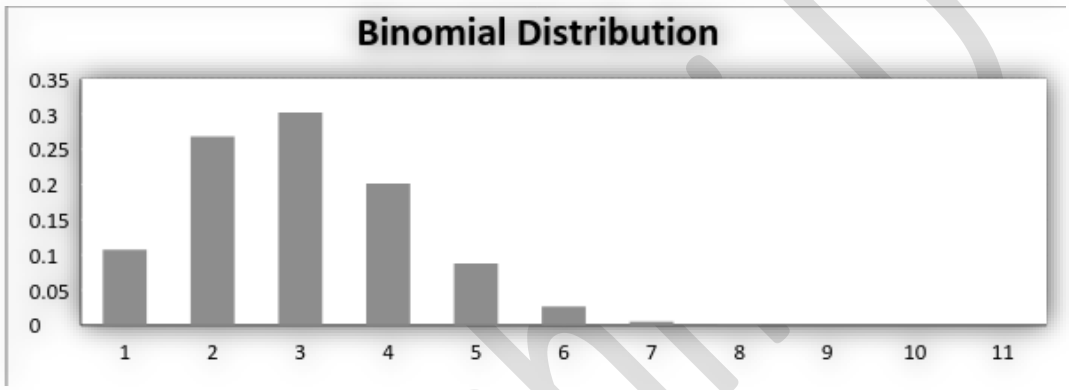
A total number of n identical trials are conducted.

The probability of success and failure is same for all trials. (Trials are identical.)

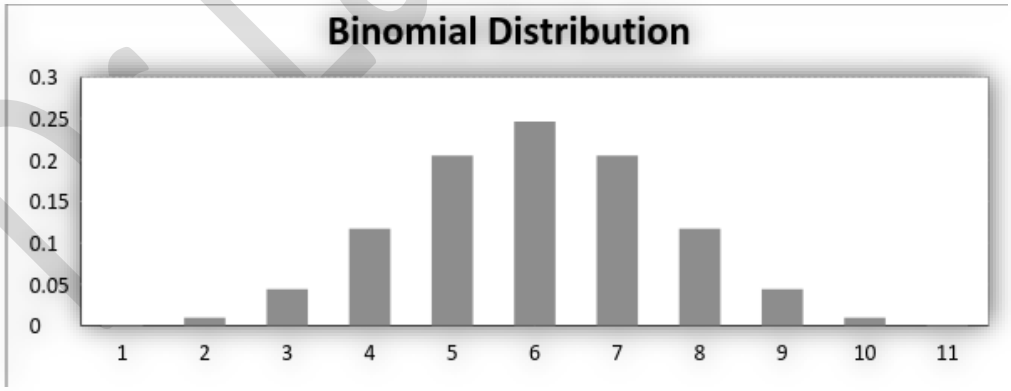
The mathematical representation of binomial distribution is given by:

$$P(x) = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

A binomial distribution graph where the probability of success does not equal the probability of failure looks like



Now, when probability of success = probability of failure, in such a situation the graph of binomial distribution looks like



The mean and variance of a binomial distribution are given by:

Mean -> μ = n*p

Variance -> Var(X) = n*p*q

Normal Distribution

Normal distribution represents the behavior **of most of the situations in the universe (That is why it's called a "normal" distribution. I guess!).** The large sum of (small) random variables often turns out to be normally distributed, contributing to its widespread application. Any distribution is known as Normal distribution if it has the following characteristics:

Section A 100 – more or less normal distribution

The mean, median and mode of the distribution coincide.

The curve of the distribution is bell-shaped and symmetrical about the line x=μ.

The total area under the curve is 1.

Exactly half of the values are to the left of the center and the other half to the right.

A normal distribution is highly different from Binomial Distribution. However, if the number of trials approaches infinity then the shapes will be quite similar.

The PDF of a random variable X following a normal distribution is given by:

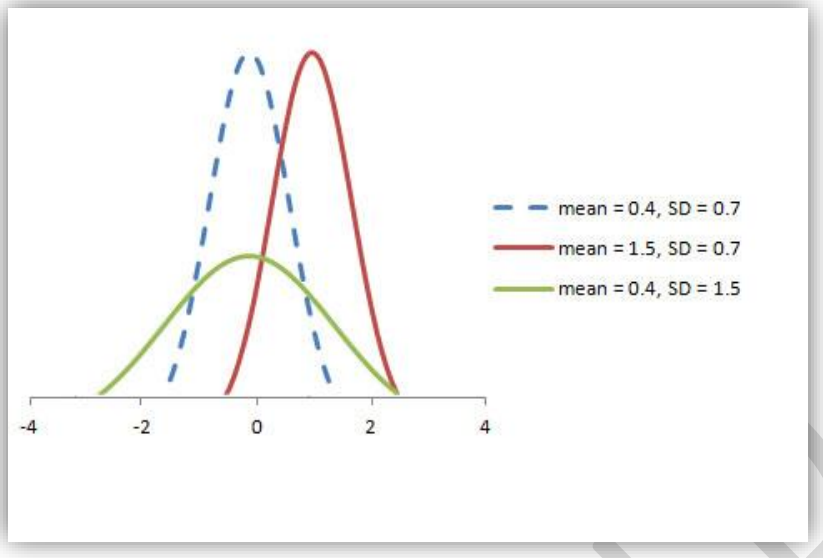
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\{-\frac{1}{2}(\frac{x-u}{\sigma})^2\}} \qquad for -\infty < x < \infty.$$

The mean and variance of a random variable X which is said to be normally distributed is given by:

Mean -> E(X) = μ

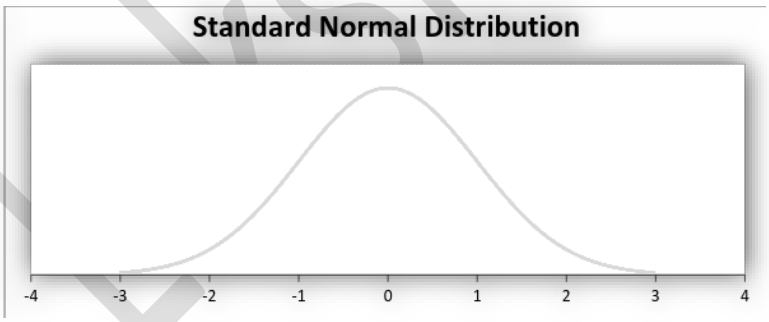
Variance -> Var(X) = σ^2

Here, μ (mean) and σ (standard deviation) are the parameters.
The graph of a random variable X ~ N (μ, σ) is shown below.



A standard normal distribution is defined as the distribution with mean 0 and standard deviation 1. For such a case, the PDF becomes:

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2} \qquad for\ -\infty < x < \infty$$



Area under the curve is 1.

Central Limit Theorem (Samples should be more)

Application: Correlation/Regression (Normal Distribution)

Confidence intervals measure the degree of uncertainty or certainty in a sampling method.

A tight interval at 95% or higher confidence is ideal.

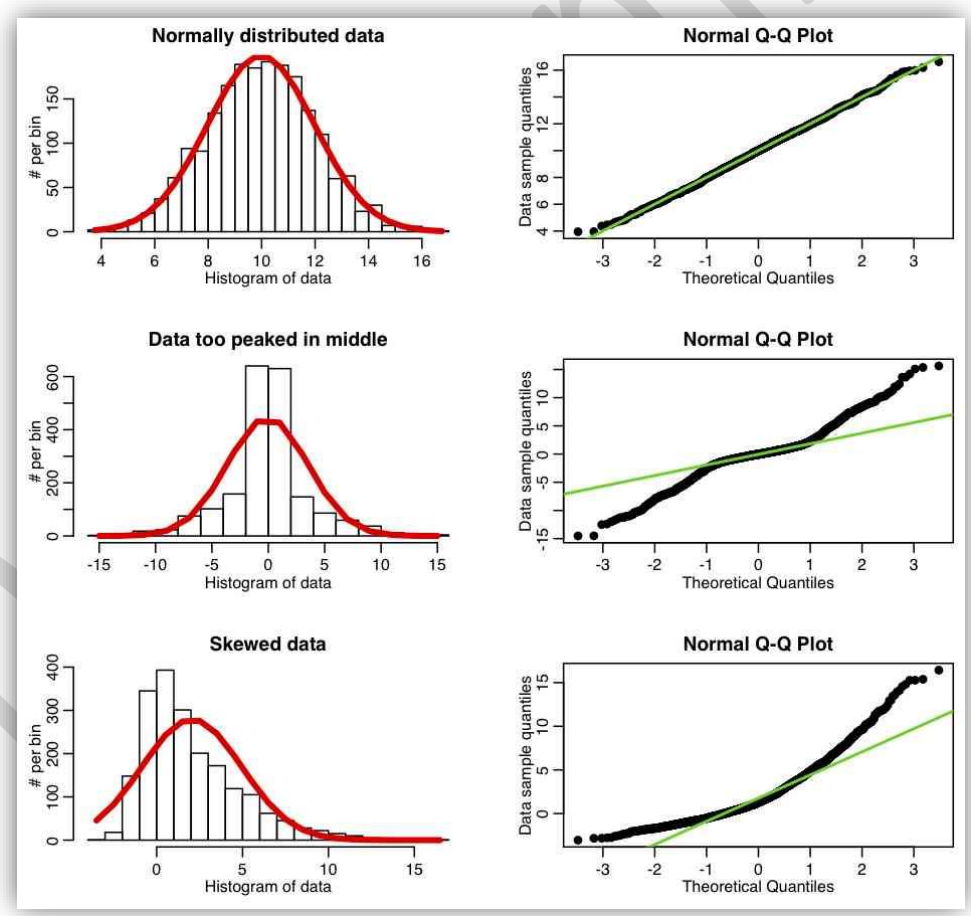
Suppose that you randomly sample light bulbs and measure the burn time. Minitab calculates that the 95% confidence interval is 1230 – 1265 hours. The confidence interval indicates that you can be 95% confident that the mean for the entire population of light bulbs falls within this range.

Confidence intervals only assess sampling error in relation to the parameter of interest. (Sampling error is simply the error inherent when trying to estimate the characteristic of an entire population from a sample.) Consequently, you should be aware of these important considerations:

As you increase the sample size, the sampling error decreases and the intervals become narrower. If you could increase the sample size to equal the population, there would be no sampling error. In this case, the confidence interval would have a width of zero and be equal to the true population parameter.

Confidence intervals only tell you about the parameter of interest and nothing about the distribution of individual values.

Accuracy 95% ➔ Reliability



Confidence Interval

Confidence Interval for μ when σ is known before hand

To gain further insight into μ, we surround the point estimate with a **margin of error**:

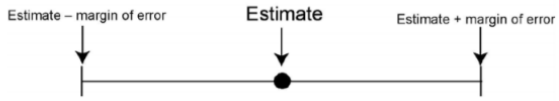
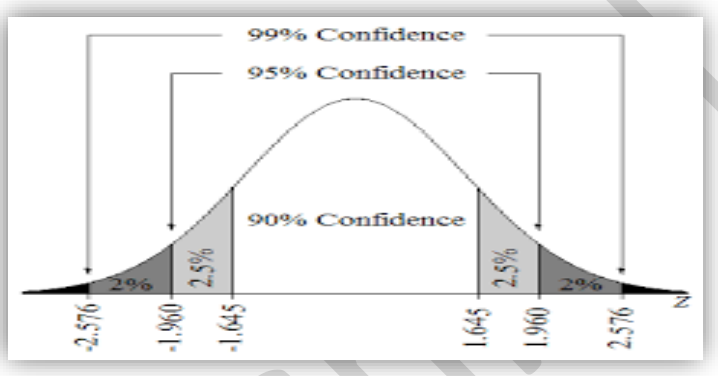


Fig: confidence-interval.ai

This forms a **confidence interval (CI)**. The lower end of the confidence interval is the **lower confidence limit (LCL)**. The upper end is the **upper confidence limit (UCL)**.



(1−α)100%	α	$z_{1-\alpha/2}$
90%	.10	1.64
95%	.05	1.96
99%	.01	2.58

Poisson Distribution

Suppose you work at a call center, approximately how many calls do you get in a day? It can be any number. Now, the entire number of calls at a call center in a day is modeled by Poisson distribution. Some more examples are

The number of emergency calls recorded at a hospital in a day.

The number of thefts reported in an area on a day.

The number of customers arriving at a salon in an hour.

The number of suicides reported in a particular city.

The number of printing errors at each page of the book.

You can now think of many examples following the same course. Poisson Distribution is applicable in situations where events occur at random points of time and space wherein our interest lies only in the number of occurrences of the event.

A distribution is called Poisson distribution when the following assumptions are valid:

1. Any successful event should not influence the outcome of another successful event.
2. The probability of success over a short interval must equal the probability of success over a longer interval.
3. The probability of success in an interval approaches zero as the interval becomes smaller.

Now, if any distribution validates the above assumptions then it is a Poisson distribution. Some notations used in Poisson distribution are:

λ is the rate at which an event occurs,

t is the length of a time interval,

And X is the number of events in that time interval.

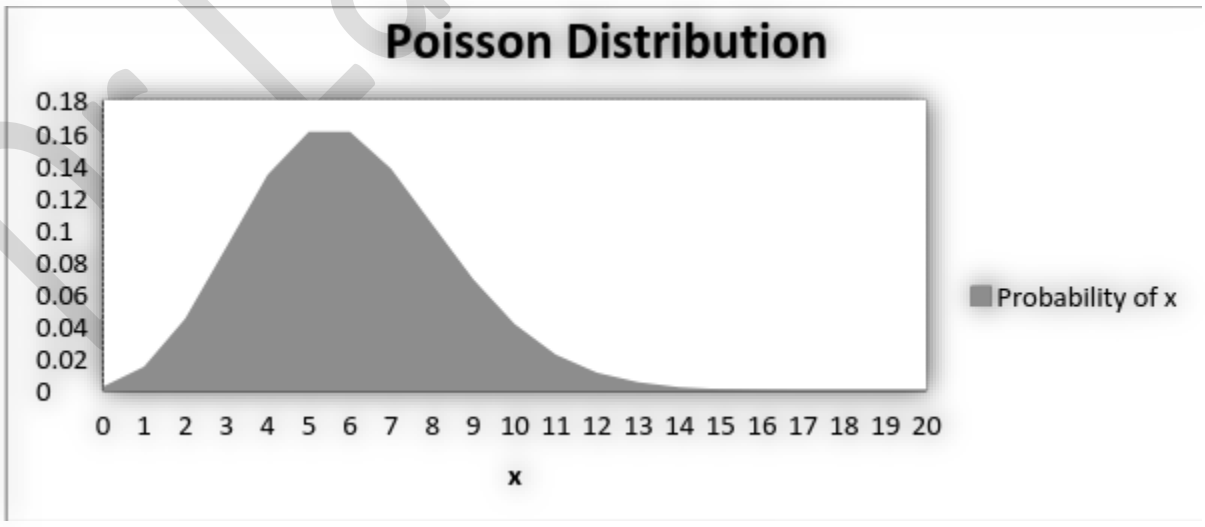
Here, X is called a Poisson Random Variable and the probability distribution of X is called Poisson distribution.

Let μ denote the mean number of events in an interval of length t. Then, μ = λ*t.

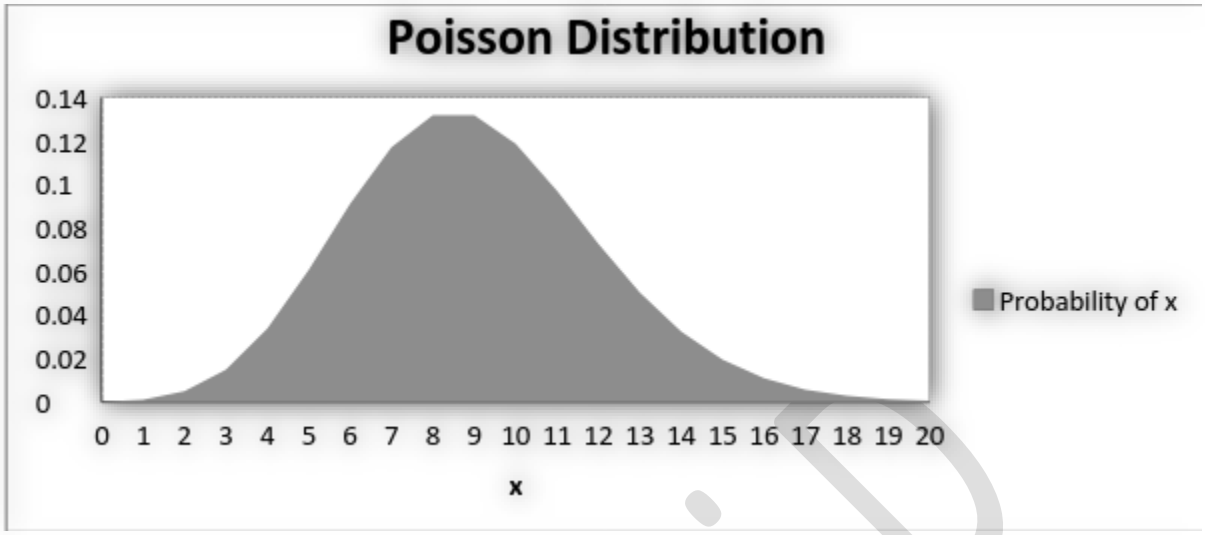
The PMF of X following a Poisson distribution is given by:

$$P(X = x) = e^{-\mu}\frac{\mu^x}{x!} \qquad for\ x = 0, 1, 2, \ldots..$$

The mean μ is the parameter of this distribution. μ is also defined as the λ times length of that interval. The graph of a Poisson distribution is shown below:



The graph shown below illustrates the shift in the curve due to increase in mean.



It is perceptible that as the mean increases, the curve shifts to the right.

The mean and variance of X following a Poisson distribution:

Mean -> E(X) = μ
Variance -> Var(X) = μ

Exponential Distribution

Let's consider the call center example one more time. What about the interval of time between the calls? Here, exponential distribution comes to our rescue. Exponential distribution models the interval of time between the calls.

Other examples are:

1. Length of time between metro arrivals,
2. The life of an Air Conditioner

Exponential distribution is widely used for survival analysis. From the expected life of a machine to the expected life of a human, exponential distribution successfully delivers the result.

A random variable X is said to have an exponential distribution with PDF:

$f(x) = \{ \lambda e^{-\lambda x}, \quad x \geq 0$

and parameter $\lambda > 0$ which is also called the rate.

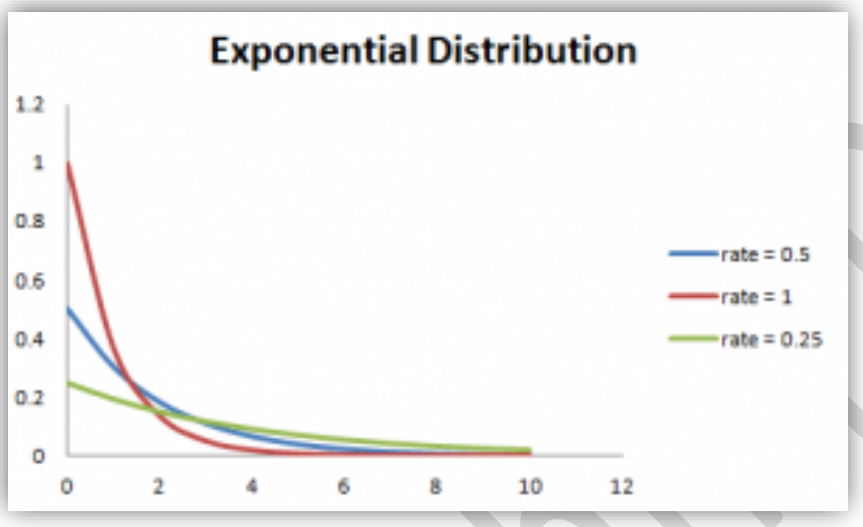
For survival analysis, λ is called the failure rate of a device at any time t, given that it has survived up to t.

Mean and Variance of a random variable X following an exponential distribution:

Mean -> $E(X) = 1/\lambda$

Variance -> $Var(X) = (1/\lambda)^2$

Also, the greater the rate, the faster the curve drops and the lower the rate, flatter the curve. This is explained better with the graph shown below.



To ease the computation, there are some formulas given below.
$P\{X \leq x\} = 1 - e^{-\lambda x}$, corresponds to the area under the density curve to the left of x.

$P\{X > x\} = e^{-\lambda x}$, corresponds to the area under the density curve to the right of x.

$P\{x1 < X \leq x2\} = e^{-\lambda x1} - e^{-\lambda x2}$, corresponds to the area under the density curve between x1 and x2.

Joint Distribution

Example for Joint Distribution is Cross Tabulation

Suppose you are a teacher at a university. After checking assignments for a week, you graded all the students. You gave these graded papers to a data entry guy in the university and tell him to create a spreadsheet containing the grades of all the students. But the guy only stores the grades and not the corresponding students.

S. No.	Scores
1	25
2	27
3	38
4	42
5	
6	16
7	35
8	46
9	48
10	31

He made another blunder, he missed a couple of entries in a hurry and we have no idea whose grades are missing. Let's find a way to solve this.

One way is that you visualize the grades and see if you can find a trend in the data.

Sampling and Estimation Methods Overview

Random Variable

Discrete Random Variable	Continuous Random Variable
Discrete Probability Distribution	Continuous Probability Distribution
Normal Distribution	Binomial Distribution

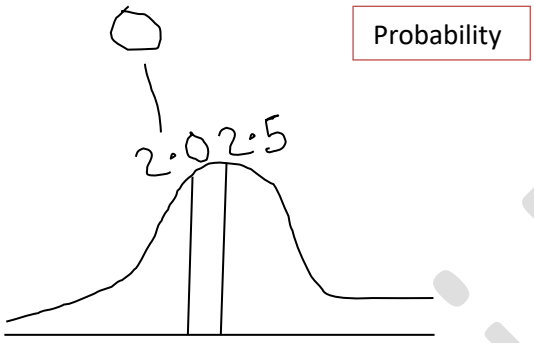
The amount of **"Rain Fall"** in a particular area is an example for random variable.

X = "The amount of Rain fall"

Let us consider the mean value is 2.5.

What is the probability of 'rain fall' for X =2.01mm?
The random variable associate with a single value is always zero.

Lesser values of the mean will occupy left of the mean, whereas, greater value will occupy right of the mean. The area at 2.0 is zero.



So always, we need to compute within the range.

P(2.01 < X < 2.5)

P(X <= 2.01) - **Calculate R, EXCEL**

P(X >= 2.01) ➔ 1- P(X <= 2.01)

Example: Find the probability, that a normally distributed random variable (Marks of students) has a mean **of μ = 60 and a standard deviation of σ =10** and we want to find the probability that **X is less than 70** (It is the mark of randomly selected student). Compute P(X<=70).

Excel normdist(X, mean, standard deviation, 1) - The Excel NORMDIST function calculates the Normal Probability Density Function or the Cumulative Normal Distribution.

R- Code pnorm(X, mean, standard deviation)

pnorm(70, 60, 10) = 0.84 (84%)

Similarly to find P(X>70). We need to do (1- Compute P(X<=70)).

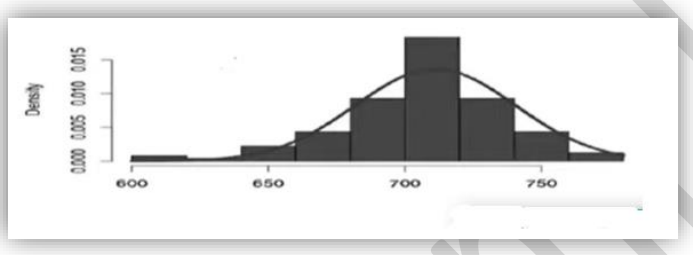
Direction of the distribution we need to consider from left to right always.

In Class Exercise

Problem 1

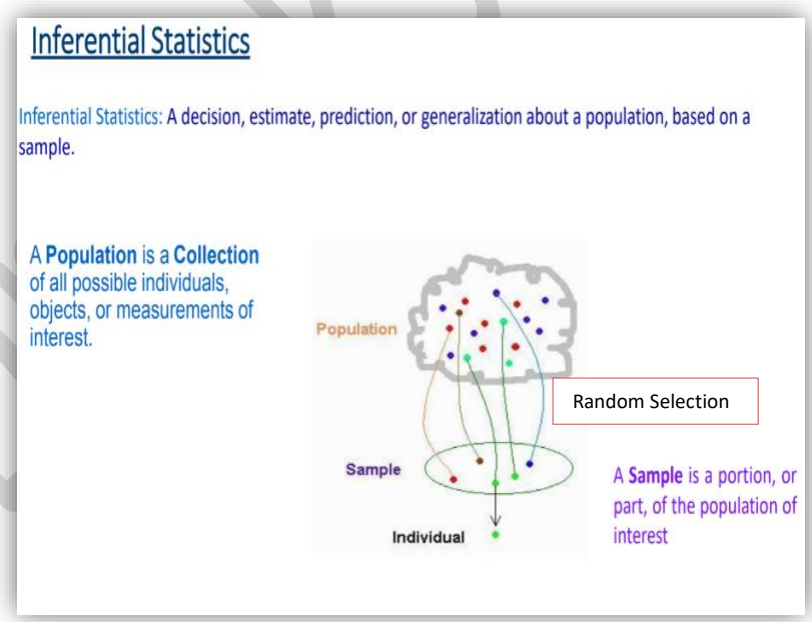
Suppose GMAT scores can be reasonably modeled using a normal distribution with **mean =711** and **standard deviation with 29.** p, q = 1-p

What is P(X<=680)? = 0.142543 **P(X > 730)? = 1- P(X < 730)? =** 0.256



Excel normdist(X, mean, standard deviation, 1) - The Excel NORMDIST function calculates the Normal Probability Density Function or the Cumulative Normal Distribution.

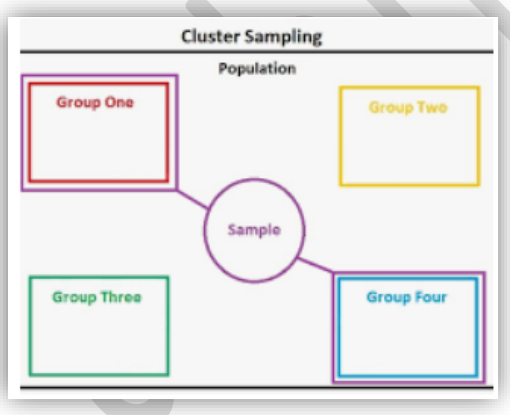
R- Code pnorm(X, mean, standard deviation)

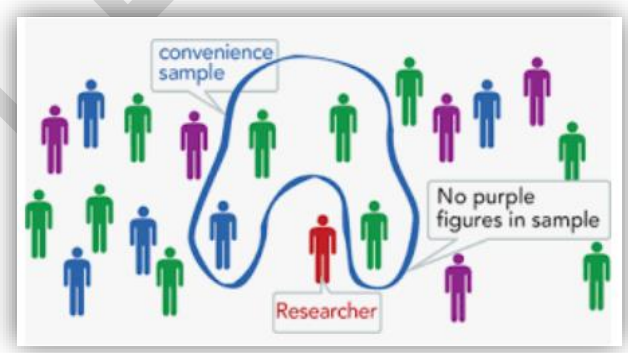


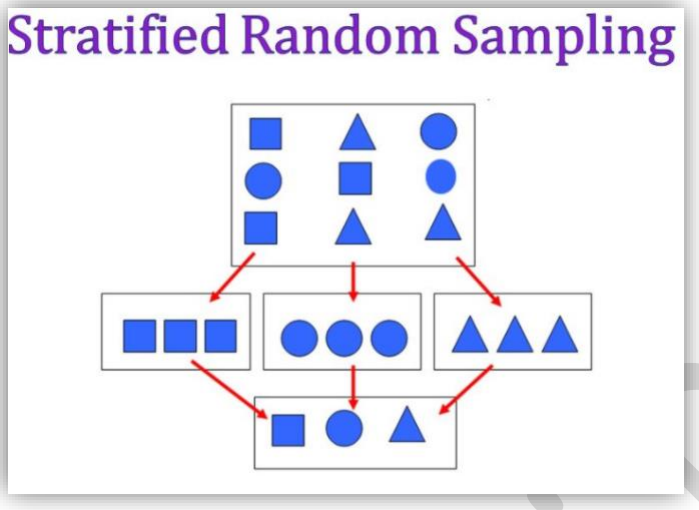
- **Sampling and Estimation methods**
- **Populations & Samples**

- **Population Mean (μ) & Sample Mean** $\bar{x}$
- **Standard Deviation: Population (σ) & Sample (s)**
- **Variance: Population (σ^2) & Sample (s^2)**
- **Proportion: Population (π) & Sample (p)**

Types of Sampling







Quota Sample

Stratified Random Sampling

- When the population under study is heterogeneous in nature then stratified random sampling is more appropriate as compared to other sampling methods to draw the representative sample of the population.
- In this method, at first population is divided into the number of subpopulation and these subpopulations are called strata.
- As far as possible make the units within the strata homogenous
- The size of each stratum may or may not be equal.
- Then sample units will be selected from each stratum independently using simple random sampling.
- This is also called restricted random sampling.
- **The stratification of the population could be done by ecological regions, development regions, rural/urban, sex, age, caste/ethnicity etc.**

Sampling Properties

Unbiased: Each unit has equal chance of being chosen in the sample.

Independent: Selection of one unit has no influence on selection of the other units.

"Simple Random Sampling" is a golden standard against all other sampling methods.

Sampling Frame

Does the sampling frame represent the population?

A sampling frame **is a list of all the items in your population.**

The available list may differ from the desired list.

Sometimes, no complete sampling frame exists.

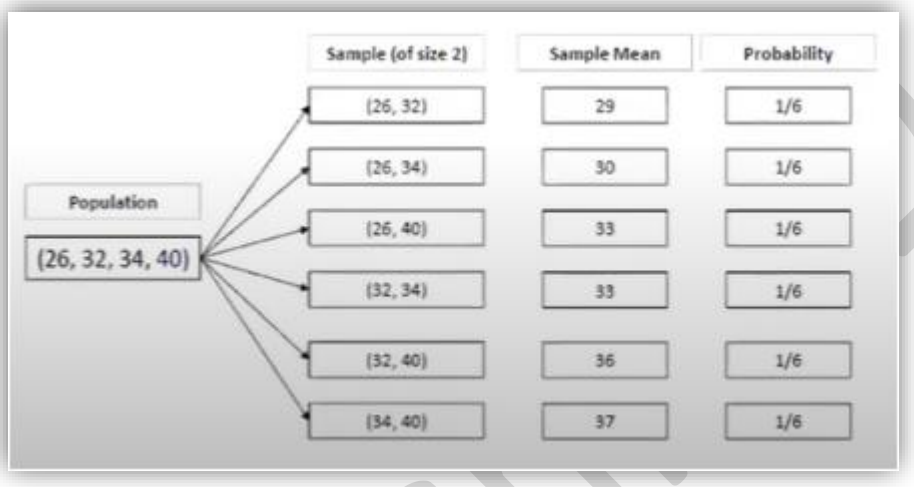
Sampling Variation

Sample mean varies from one sample to another

Sample mean can be different from the population mean

Sample mean is random variable

Sample mean near to population mean



Statistical Inference

The two common forms of statistical inference are:

- **Estimation**
- **Null hypothesis tests of significance (NHTS)**

There are two forms of estimation:

- **Point estimation (maximally likely value for parameter) Ex. mean**
- **Interval estimation (also called confidence interval for parameter) [CI]**

Confidence Interval

[73] - Confidence 0%

[50 60] - 30% ➔ Narrow Confidence

[30 70] - 60%

[20 80] - 80% [Right choice]

[10 90] - 90%

[0 100] - 100%

Central Limit Theorem – Normal Distribution [most statistical tools are working / designed based on the underlying principle called normal distribution] – Natural events are resembling the normal distribution

Animation Link

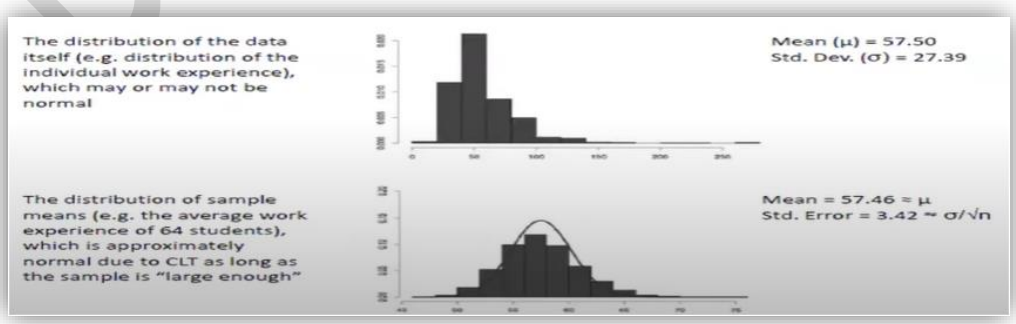
https://www.geogebra.org/m/zshvnvuj

https://www.geogebra.org/t/math

The distribution of the sample mean

- **Will be normal when the distribution of data in the population is normal.**
- **Will be approximately normal even if the distribution of data in the population is not normal, if the sample size is "larger".**
- **thumb rule to check 'Central Limit Theorem'**

n >= 10 * (*skewness*)^2

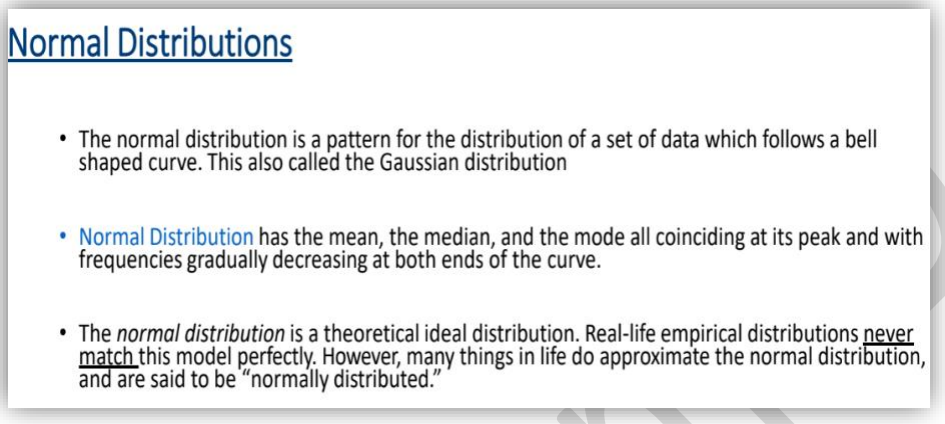


Excel Demonstration

Set of Samples with rand()

Compute average()

Normal Distribution Curve / Plot (Using Histogram)



Confidence Interval

Demonstration with task assigned and the timeline

Exactly 70 hours – 0% confidence

Between [90 110] 40%

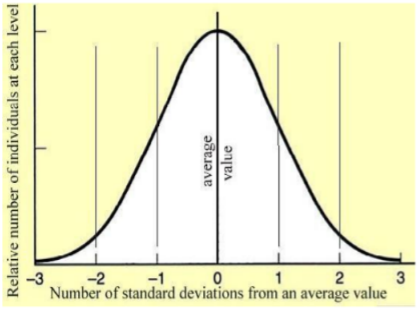
[90 150] 70%

[0 300] 100%

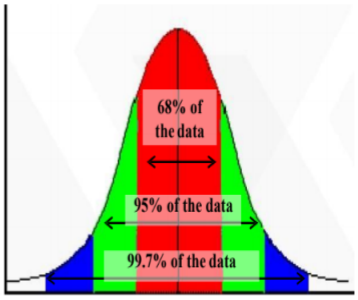
However, the interval is increased there is no chance for the earlier prediction.

The Bell Shaped Curve

- The bell shaped curve has the following characteristics:
  - The curve is concentrated in the center and decreases on either side.
  - The bell shaped curve is symmetric and Unimodal
  - The curve extends to + / - infinity
  - Area under the curve = 1



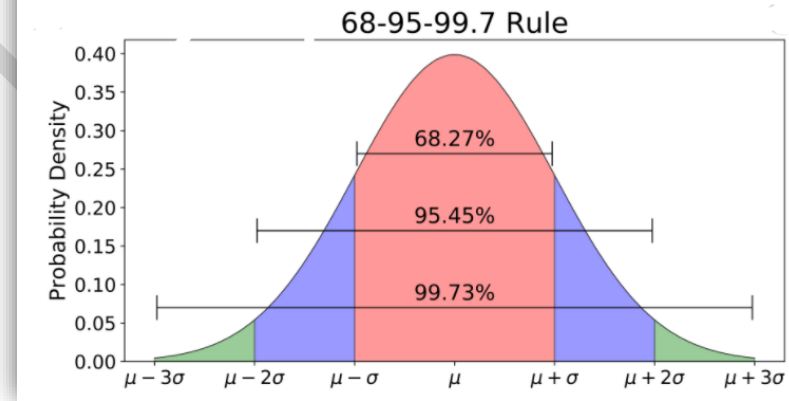
68-95-99.7 Rule



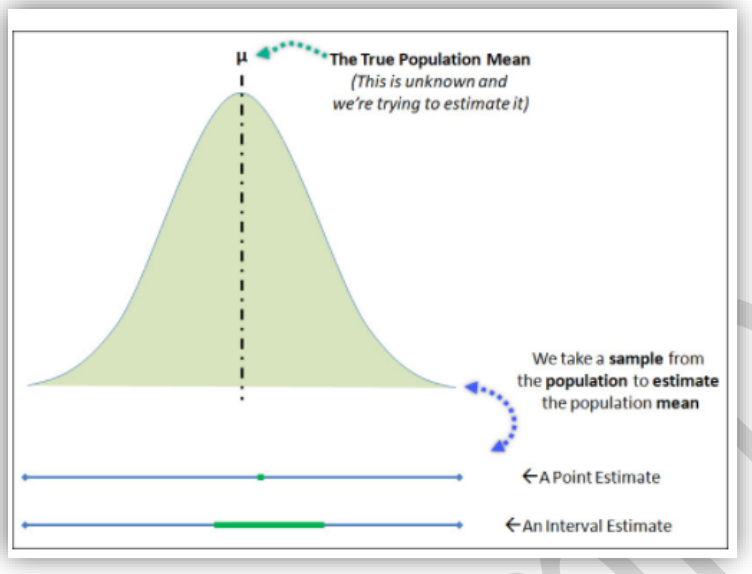
The empirical rule states that for a normal distribution:

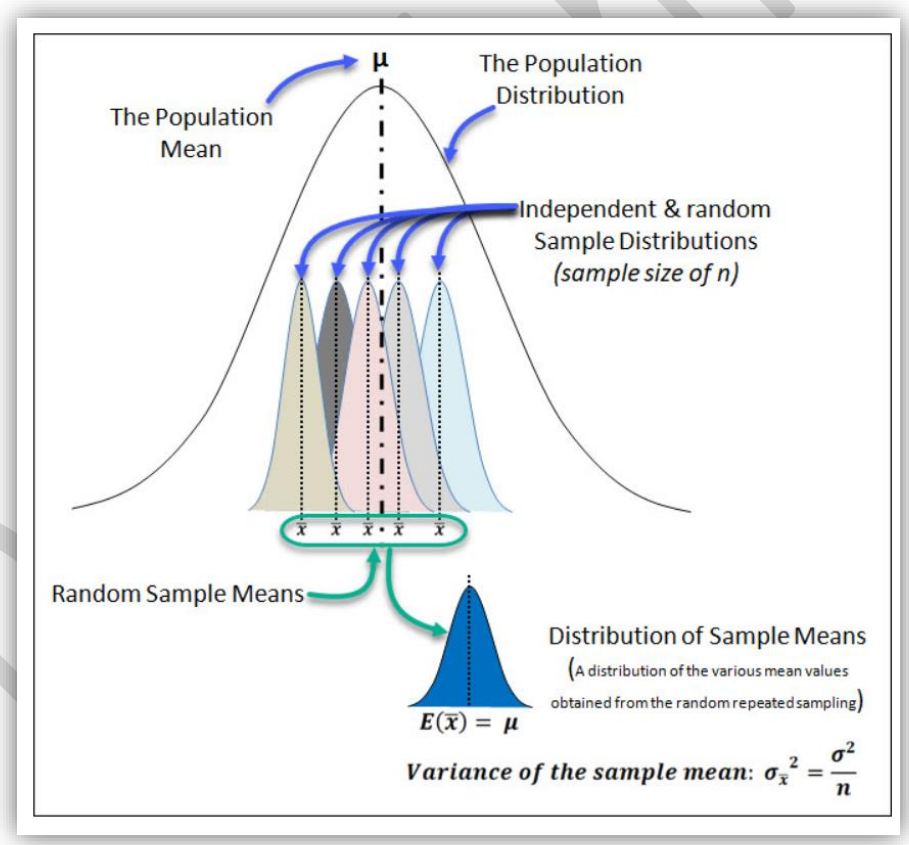
- 68% of the data will fall within 1 SD of mean
- 95% of the data will fall within 2 SD's of the mean
- Almost all (99.7%) of the data will fall within 3 SD's of the mean

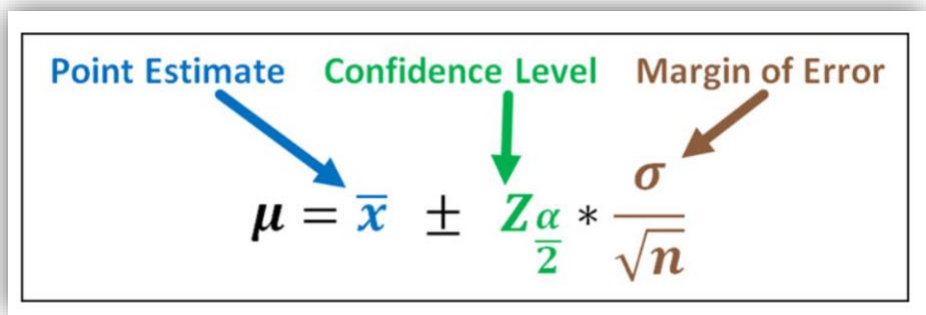
$(1-\alpha)100\%$	α	$z_{1-\alpha/2}$
90%	.10	1.64
95%	.05	1.96
99%	.01	2.58

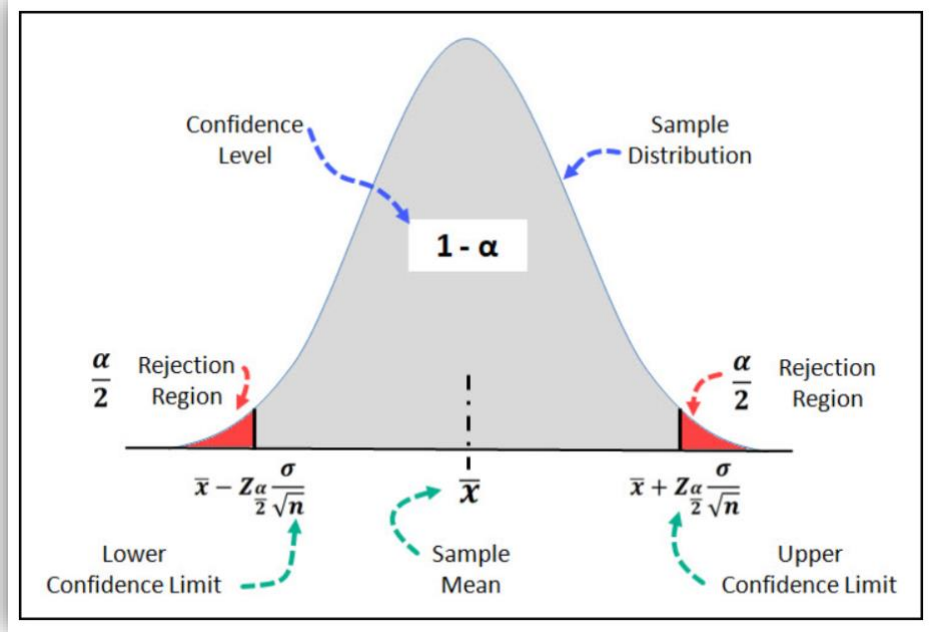


Confidence interval for the population mean





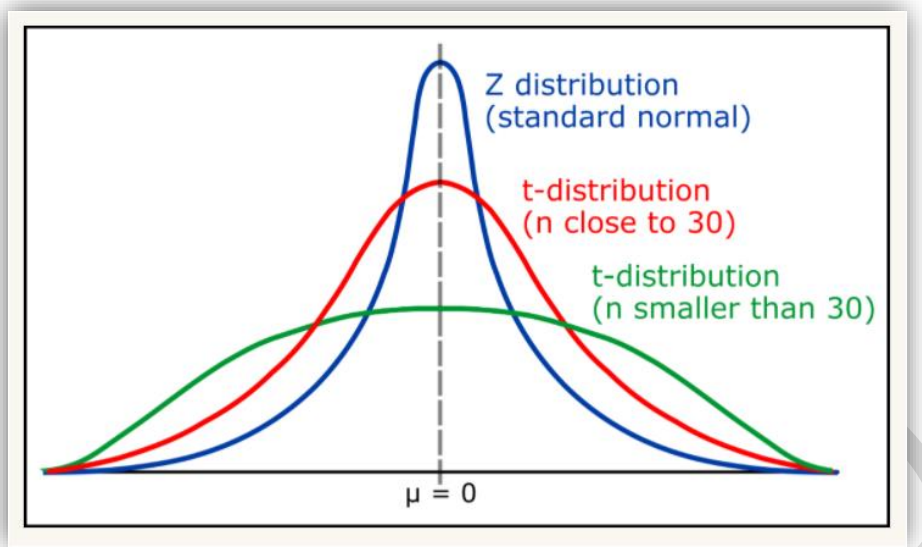




$$\textbf{Interval Estimate of Population Mean (known variance)} : \bar{x} \pm Z\frac{\alpha}{2} * \frac{\sigma}{\sqrt{n}}$$

When the population variance is unknown, you use the t-distribution (t-score) and the sample variance to create your interval estimate using the following equation:

$$\textbf{Interval Estimate of Population Mean (unknown variance)} : \bar{x} \pm t\frac{\alpha}{2} * \frac{s}{\sqrt{n}}$$



Which test we need to use z-test or t-test?

Example of Interval Estimate of Population Mean with Known Variance

You've sampled 40 units from the latest production lot to measure the weight of the product, and the sample mean is 10.40 lbs. If the population standard deviation is known to be 0.60 lbs, calculate the 95% confidence interval.

Ok, let's see what we know after reading the question:

n = 40, σ = 0.60 lbs, α = 0.05, x-bar = 10.4 lbs.

Before we can plug this into our equation we need to find the Z-score associated with the 95% confidence interval.

If we look that up in the NIST Z-Table, we find Z = 1.96.

The Z-score of 1.96 is associated with an area under the curve of 0.475.

This is because the normal distribution is two-sided and the alpha risk associated with one side of the distribution is 0.500 – 0.025 = 0.475.

$$Interval\ Estimate\ of\ Population\ Mean\ (known\ variance)\ : \bar{x} \pm Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

$$Interval\ Estimate : 10.4 \pm 1.96 * \frac{0.60}{\sqrt{40}}$$

$$Interval\ Estimate : 10.4 \pm 0.186$$

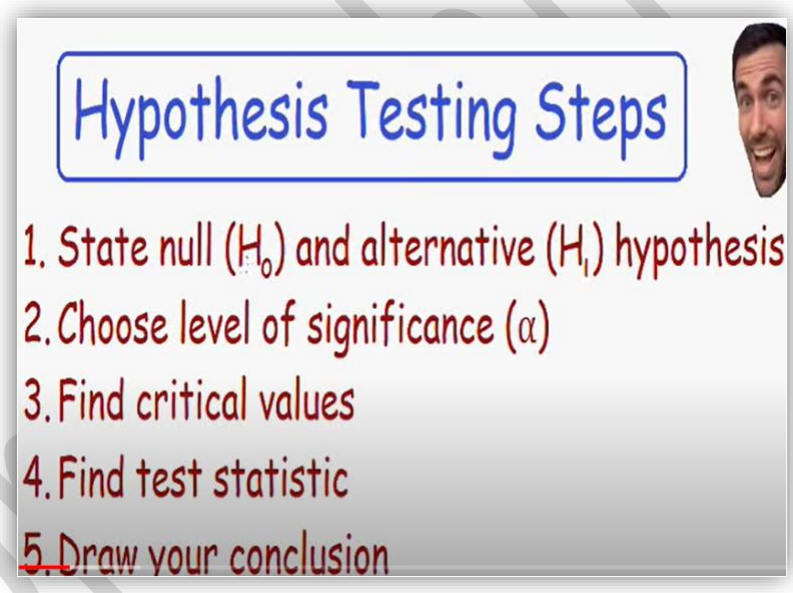
$$95\%\ Confidence\ Interval : 10.21 - 10.59$$

The average IQ of the adult population is 100.
A researcher believes the average IQ of adults is lower.
A random sample of 5 adults are tested and scored
69, 79, 89, 99, 109. (s.d. = 15.81)
Is there enough evidence to suggest the average IQ is lower?

Hypothesis

H0: $\mu = 100$ (There is no difference/Discourage)

H1: $\mu < 100$ (Encourage)



one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05
df							
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571

Significance Level ➔ 0.05

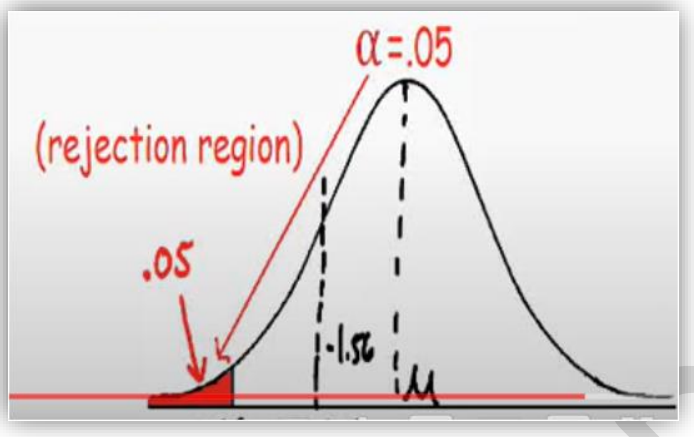
Critical Value ➔ 2.132

t = x - μ / ***(s/SQRT(n))***

$$\frac{89-100}{15.81/\sqrt{5}} = -1.56$$

test statistic

The obtained t value is less than the p-value; hence we need to accept the null hypothesis. The conclusion is average IQ of the given adult set also fall into the population IQ level.



Conclusion:

The obtained t-value fits with the confidence interval. Hence the sample adults also possess the same IQ level of the population.

Credit Card Launch

A university with 10 000 alumni is thinking of offering a new affinity credit card to its alumni.

Profitability of the card depends on the average balance maintained by the cardholders.

A market research campaign is launched, in which about 140 alumni accept the card in a pilot launch.

Average balance maintained by these is $1990 and the **standard deviation is $2833**. Assume that the **population standard deviation is $2500 from previous launches**.

What can we say about the average balance that will be held after a full-fledged market launch?

Population: 1000, 000
Sample Size: 140
Mean: 1990
Sample Standard Deviation: 2833
Population Standard Deviation: 2500
Confidence Interval 95%

Based on the sample mean, **we should not take the decision, because it has involved with uncertainty**.

$$\bar{x} \pm Z\frac{\alpha}{2} * \frac{\sigma}{\sqrt{n}}$$

= 1990 – (1.96) * ((2500/sqrt(140)) 1990+(1.960) * ((2500/sqrt(140))

= [1990 – (1.96) * 211.29 1990 + (1.960) * 211.29]

= [1575 2404] 95% ➔ [$1575 $2404]

Confidence Level		z
0.70	70%	1.04
0.75	75%	1.15
0.80	80%	1.28
0.85	85%	1.44
0.90	90%	1.645
0.92	92%	1.75
0.95	95%	1.96
0.96	96%	2.05
0.98	98%	2.33
0.99	99%	2.58

R (Packages) and Excel search for commands related to Confidence Interval.

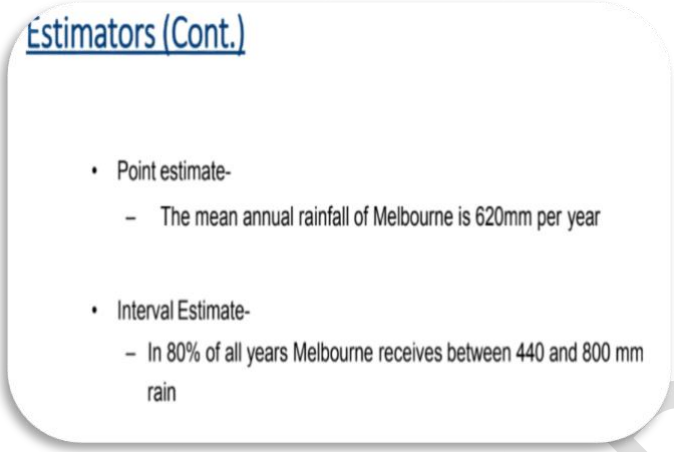
Compute the probability pnorm() for z-test, pt() for t-test in R

Hypothesis Testing

It is another inferential statistics. Let us consider the below example. Now we need to compare whether these two processes are statistically different or same.

First step we need to find out descriptive statistics, to have an overall understanding such as mean and standard deviation.

With the hypothesis testing only we can conclude about the sample and its inferences.



Defining Hypothesis

H0, H_0 (Null Hypothesis) $\mu_A = \mu_B$ (**No Change**)
HA, H_A (Alternative Hypothesis) μ_A not equal to μ_B (**There is a considerable change**)

$Z = \bar{x} - \mu / \sigma$ ***- If we know population standard deviation and sample size is more than 30, we need to calculate z-test***
$t = \bar{x} - \mu /$ ***(s/SQRT(n)) - Otherwise t-test***

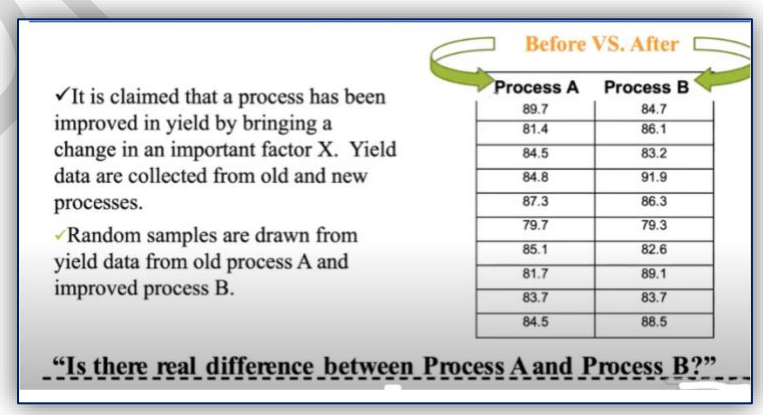
p-value, alpha-value

Probability value or p-value

Fix alpha value as 0.05 [Suitable alpha value needs to be decided by experts]

Compare alpha with p-value

If (p-Value <Alpha) Reject H0 (The claim of both the processes are equal)



Pharma A Pharma B (Which drug is better whether drug A & drug B)

Covid vaccination

Sample A (200) Sample B (200)

3 4

4 5

H0: There is no significant difference between drug A and drug B

HA: There is a significant difference between these two drugs/ drug A and drug B

Hypothesis Testing (cont..)

In a Test Procedure, to start with, a hypothesis is made.

The validity of the hypothesis is tested.

If the hypothesis is found to be true, it is accepted.

If it is found to be untrue, it is rejected.

The hypothesis which is being tested for possible rejection is called null hypothesis

Null hypothesis is denoted by H_0

The hypothesis which is accepted when null hypothesis is rejected is called Alternate Hypothesis H_a

Hypothesis Testing (cont..)

The alternative hypothesis is often the interesting one –and often the one that someone sets out to prove.

Ex. H_o: The drug works –it has a real effect.

H_a: The drug doesn't work - Any effect you saw was due to chance.

Think of "null" as "there's no real effect," and "alternative" as "other than null."

As we now know how to apply confidence intervals to population means, the next logical step is to infer about values outside this range…

We will learn an important technique known as hypothesis testing, where we hypothesise the true population parameter, and then use the confidence interval logic to test the probability of the hypothesis

Hypothesis Testing (cont..)

Hypothesis tests consist of the following steps:

- Null hypothesis
- Alternative hypothesis
- Confidence level
- Decision Rule
- Test statistic
- Decision

An example of status quo is the economic conditions of a particular class at a particular period of history.

Example: Supermarket Loyalty Program

- A supermarket plans to launch a loyalty program if it results in an average spending per shopper of more than $120 per week
- A random sample of 80 shoppers enrolled in the pilot program spent an average of $130 in a week with a standard deviation of $40
- Should the loyalty program be launched?

$\bar{X} = 130$

s = 40

n= 80

$\mu = 120$

Two Tail or One Tail

> or < one tail

= not = two tail

One Tail ➔ When to use? ➔ less than or greater than

Two tail ➔ when to use? ➔ Equal to or not equal to

H0 ➔ Don't take any action [discourages] $\mu <= 120$

HA ➔ You can take action if $\mu > 120$

We need to go for t-test 130-120/(40/SQRT(80)) =

pt(t-test, degrees of freedom n-1) ➔ pt(2.2., 79) = R gives always from left to right

Always check for alternate hypothesis to decide whether right or left side of the mean

Since it is greater than, we need to calculate

1-pt(2.2., 79) = 0.015 (P value)

Alpha is 0.05

Now compare both of them

p-value is less than Alpha, so we need to reject null hypothesis.

Hence final conclusion is we can launch the loyalty scheme to all the customers based on alternative hypothesis.

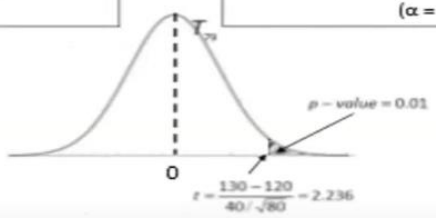
Fixing H0 and HA need to be cautious. H0 always discourages, HA always encourages [remember this view point's] then decide H0, HA.

Right-tailed hypothesis test (sample mean of 130)

Null Hypothesis: The additional spending is less than or equal to $120
H0: $\mu \leq \mu_0$ (120)

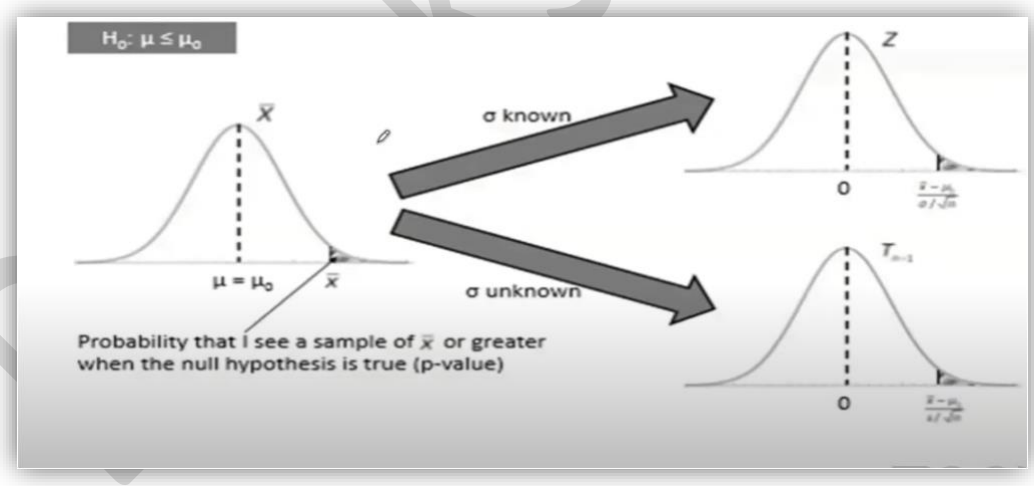
Acceptable level of type-I error is 0.05

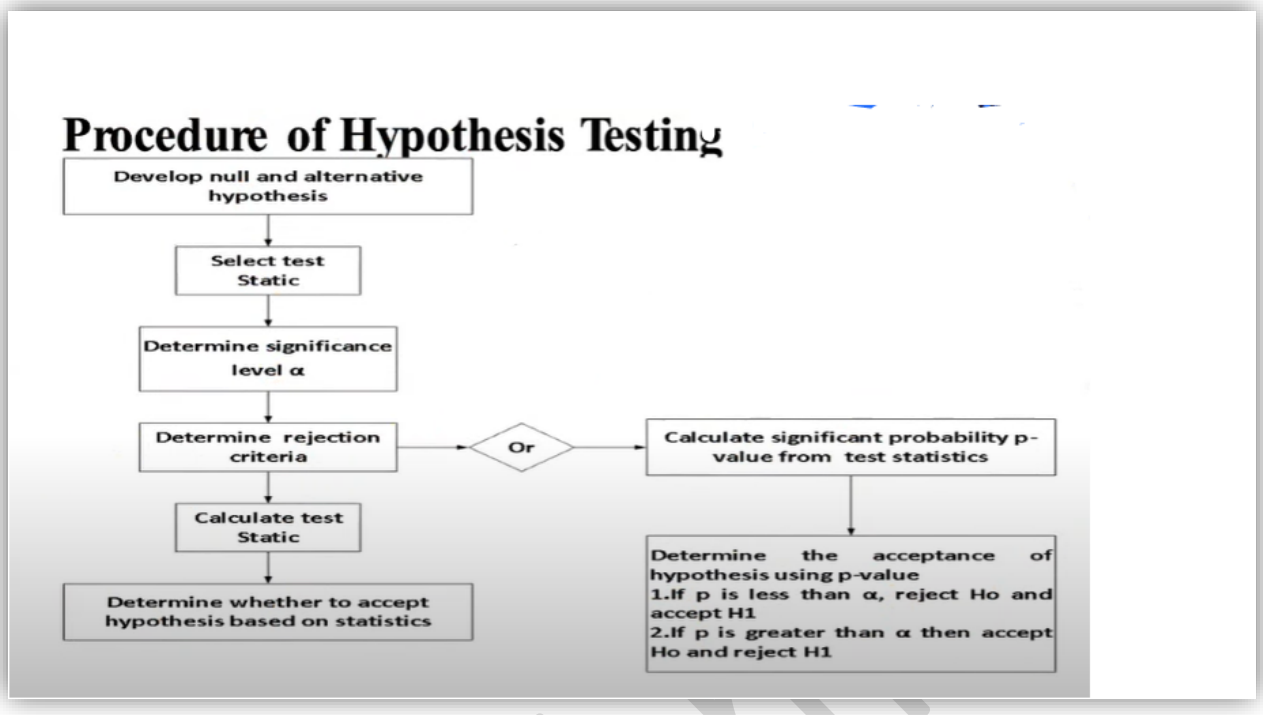
($\alpha = 0.05$)



If I reject my null hypothesis that $\mu \leq 120$, I will be wrong with probability 0.01

Launch the loyalty card





Equal to or Not equal to ➔ Two tail (normal Distribution)

Greater or Lesser ➔ One tail

Variants of t-test

1-sample t-test Marks of a class We need to compare with standard value	**1-tail t-test** **This is used to test either one side of the tail in the normal distribution** **H0: <** **HA: >** **Relationship** **One p-value**
2-sample t-test Class A | Class B	**2-tail t-test** **H0: =** **HA: <>**

		This is used to test on both sides of the tail in the normal distribution **Final p-value is p1 + p2 Combining on both sides**

Process control at a call center

Performance of a call center is monitored by the average call duration.

Data from 20 days shows that on the days when the process runs normally

μ = 4 minutes, s= 3 minutes

Cannot monitor each and every call due to limited resources; so randomly sampled 50 calls per day

Day	Mean Call Duration
1	3.7
2	4.1
3	3.5
4	4.2
5	3.9
6	4.1
7	4.2
8	3.8
9	3.7
10	4.6
11	3.7
12	4.6
13	4.0
14	4.2
15	3.8
16	4.4
17	5.3
18	6.1
19	7.2
20	6.5

Average of each day for sample size of 50

We need to verify whether each day the process is under control of out of control

Sample size n=50

Days are 20

We can consider for example 10^{th} day whether the process is under control or not?

H0: μ = 4 (Will not encourage)

HA: μ ≠ 4

If it is less than or more than 4 minutes are not to be considered.

Here when the process is out of control, we need to take decisions, H0 will not encourage for the decisions, whereas, HA will allow us to take decision.

The discussed example suits for which category?

It is related to one sample two tail test.

It is a 2 tail t-test

t = xbar - μ/ (s/ sqrt(n))

= 4.6 – 4/(3/sqrt(50) = 1.4

Now let us calculate p-value (P1 + P2)

=2*(1- pt(1.4, 49))

=0.16

Now let us form the decision rule α=0.05

Is (0.16<0.05) p-value is less than alpha

No

Recommendation

Hence, we reject H0 or accept alternative hypothesis. This indicates the process is not the under control, so need of tracing the root cause for the delay is essential. Need to improve the process.

Hypothesis test using R

y is t.test(x,y=NULL, alternative = c("two.sided", "less", "greater"), mu=0, var.equal=FALSE, conf.level = 0.95)

x- vector data

y- optional vector data

meant for two sample test

Example

An outbreak of **Salmonella-related illness** was attributed to ice cream produced at a certain factory. Scientist measured the level of Salmonella in 9 randomly sampled batches of ice cream. The levels (in MPN/g) were

0.593 0.142 0.329 0.691 0.231 0.793 0.519 0.392 0.418

Is there evidence that the **mean level of Salmonella** in the ice cream is **greater than 0.3 MPN/g?**

https://www.statext.com/practice/OneT02.php

What type of test we need to proceed?

It is one sample one tail test.

Let μ = mean level of Salmonella in all batches of ice cream.

H_0: $\mu <= 0.3$

H_a: $\mu > 0.3$

= 0.456, S = 0.2128, n = 9

We will use the t-test since n < 30.

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

t = 0.4564-0.3/(0.213/3) = 2.2055

t table for .05 level of significance for 8 degrees of freedom is 1.860.

Reject the null hypothesis since 2.2055> 1.860.

pt(2.2055, 8)

Is (1.860<0.05)?

No

Alternative Hypothesis

Recommendation

The level of **Salmonella level** is beyond the given mean level) is greater than 3, it is necessary to take action on the producer. Or the circular need to be sent to the producers to reduce the **Salmonella level or control.**

Let mu be the mean level of Salmonella in all batches of ice cream. Here the hypothesis of interest can be expressed as:

H0: mu= 0.3

Ha: mu> 0.3

Hence, we will need to include the options alternative="greater", mu=0.3.

```
x = c(0.593, 0.142, 0.329, 0.691, 0.231, 0.793, 0.519, 0.392, 0.418)
t.test(x, alternative="greater", mu=0.3)
```

p-value = 0.029. Hence, there is moderately strong evidence that the mean Salmonella level in the ice cream is above 0.3 MPN/g.

https://www.slideshare.net/karishmasharma/hypothesis-testing-in-r

```
data: x
t = 2.2051, df = 8, p-value = 0.02927
alternative hypothesis: true mean is greater than 0.3
```

From the output we see that the p-value = 0.029. Hence, there is moderately strong evidence that the mean Salmonella level in the ice cream is above 0.3 MPN/g.

Example

Ex. 6 subjects were given a drug (treatment group) and an additional 6 subjects a placebo (control group). Their reaction time to a stimulus was measured (in *ms). We want to perform a two-sample t-test for comparing the means of the treatment and control groups.*

Control : 91, 87, 99, 77, 88, 91
Treat :101, 110, 103, 93, 99, 104

The values are reaction time of the drug

Ho: $\mu_{Control} = \mu_{Treat}$ [There is no big difference between both of the medicines]

HA: $\mu_{Control} \neq \mu_{Treat}$ [There is a difference between both of the medicine]

Two sample two tail test

Let Mu1 be the mean of the population taking medicine and Mu2 the mean of the untreated population. Here the hypothesis of interest can be expressed as:

H0: Mu1-Mu2=0
Ha: Mu1-Mu2 !=0

Here we will need to include the data for the treatment group in x and the data for the control group in y. We will also need to include the options alternative="less", mu=0.

Finally, we need to decide whether or not the standard deviations are the same in both groups.

```
> Control = c( 91, 87, 99, 77, 88, 91)
> Treat = c( 101, 110, 103, 93, 99, 104)
> t.test (Control, Treat, alternative="two.sided")
    Two Sample t-test
    data: Control and Treat
    t = -3.4456, df = 9.4797, p-value = 0.006782
```

R Output

```
        Welch Two Sample t-test

data:  Control and Treat
t = -3.4456, df = 9.4797, p-value = 0.006782
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -21.194292  -4.472375
sample estimates:
mean of x mean of y
 88.83333 101.66667
```

Refer the following link for calculating degrees of freedom

https://www.statsdirect.co.uk/help/parametric_methods/utt.htm#:~:text=This%20function%20gives%20an%20unpaired,the%20difference%20between%20the%20means.&text=%2D%20where%20x%20bar%201%20and,2%20%2D%202%20degrees%20of%20freedom.

Bank Marketing Data Set - Intelligent Targeting

Marketing campaigns are characterized by focusing on the customer needs and their overall satisfaction. Nevertheless, there are different variables that determine whether a marketing campaign will be successful or not. Some important aspects of a marketing campaign are as follows:

Segment of the Population: To which segment of the population is the marketing campaign going to address and why? This aspect of the marketing campaign is extremely important since it will tell to which part of the population should most likely receive the message of the marketing campaign.

Distribution channel to reach the customer's place: Implementing the most effective strategy in order to get the most out of this marketing campaign. What segment of the population should we address? Which instrument should we use to get our message out? (Ex: Telephones, Radio, TV, Social Media Etc.)

Promotional Strategy: This is the way the strategy is going to be implemented and how are potential clients going to be address. This should be the last part of the marketing campaign analysis since there has to be an in-depth analysis of previous campaigns (If possible) in order to learn from previous mistakes and to determine how to make the marketing campaign much more effective.

You are leading the marketing analytics team for a banking institution. There has been a revenue decline for the bank and they would like to know what actions to take. After investigation, it was found that the root cause is that their clients are not depositing as frequently as before. Term deposits allow banks to hold onto a deposit for a specific amount of time, so banks can lend more and thus make more profits. In addition, banks also hold better chance to persuade term deposit clients into buying other products such as funds or insurance to further increase their revenues.

You are provided a dataset containing details of marketing campaigns done via phone with various details for customers such as demographics, last campaign details etc. Can you help the bank to predict accurately whether the customer will subscribe to the focus product for the campaign - Term Deposit after the campaign?

Data Description

- We can see that variables 1 to 16 can be used for modelling.
- Variable '**id**' is Identifier column. It has a unique value for every sample in the dataset and cannot be used for modelling.
- Variable '**term_deposit_subscribed**' is Target/y column. It has binary values and we need to learn to predict this variable given our above 16 variables as features.

Hypothesis Generation

Simply put, a hypothesis is a possible view or assertion of an analyst about the problem he or she is working upon. It may be true or may not be true.

- Are younger customers more likely to subscribe to a term deposit as compared to old customers ?
- Are people with a higher bank balance more likely subscribe to term deposit than people with low balance ?
- Does a married person have higher chances to subscribe to a term deposit compared to a single or divorced person ?
- Does the length of the call made to the customers tell us anything about their chances of subscribing ? (One may think that if the call length is longer, the customer executive has spent a longer time, discussing details with customer, hence the customer is more likely to subscribe).

References

http://www.cqeacademy.com/cqe-body-of-knowledge/quantitative-methods-tools/point-estimates-and-confidence-intervals/

https://andyjconnelly.wordpress.com/2017/05/16/uncertainty-and-repeats/

https://crumplab.github.io/statistics/probability-sampling-and-estimation.html#idea-behind-z-scores

https://www.cfainstitute.org/en/membership/professional-development/refresher-readings/2020/sampling-estimation

https://machinelearningmastery.com/statistical-sampling-and-resampling/

https://www.statisticshowto.com/probability-and-statistics/sampling-in-statistics/

https://www.youtube.com/watch?v=1hSK_brcMfE&t=1482s&ab_channel=ExcelRSolutions-RaisingExcellence

https://www.youtube.com/watch?v=_D5iAvJhQ5Y&list=PLkqc8xRb_lICNWkMuAO4sygNWK7iRLrmM&index=5&ab_channel=ExcelRSolutions-RaisingExcellence