Statistics for DATA SCIENCE

Topics

- Why Statistics and Big data
- Types of stats
- Some vital terms in stats
- Sources and types of data and datasets
- Data objects, attributes and its types
- Descriptive Statistics outline
- Data and Histogram
- Central tendency and 3 Ms

1. Why Statistics and Big data

The significant events triggered the current meteoric growth in the use of analytical

decision making and Statistics is central to all of them.

Event 1:

- Technological developments, Revolution of Internet and Social Networks, data generated from mobile phones produce large amount of data from which insights will be shifted.
- The discovery of pattern and trends from these data for organizations will pay the way for improving profitability, understanding customer expectations so that they can gain competitive advantage in the market place.

Event 2:

- Advances in enormous computing power to effectively process and analyse massive amounts of data.
- Sophisticated and faster algorithms for solving problems.
- Data visualization for Business intelligence and artificial intelligence.

Event 3:

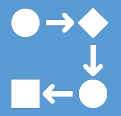
- Large data storage capability.
- Parallel and cloud computing have enabled business to solve large scale problems.

Types of Data

Qualitative data are non numeric in nature and cannot be measured.

Quantitative data are numerical in nature and can be measured and can be classified into two: discrete and continuous.

Discrete type can take only certain values, and there are discontinuities between values.



Continuous type can take any value within a specific interval.

Data objects, attributes and its types

Nominal: categories, states or “names of things”

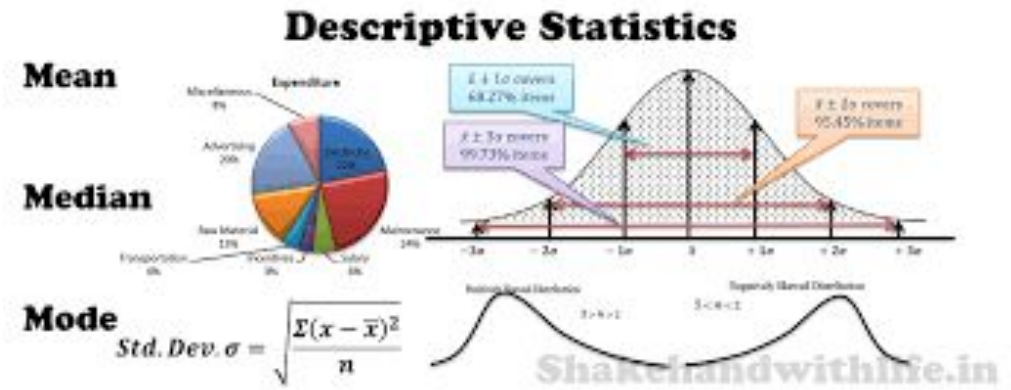
- Hair_color = {black, brown, grey}
- marital status, occupation, ID

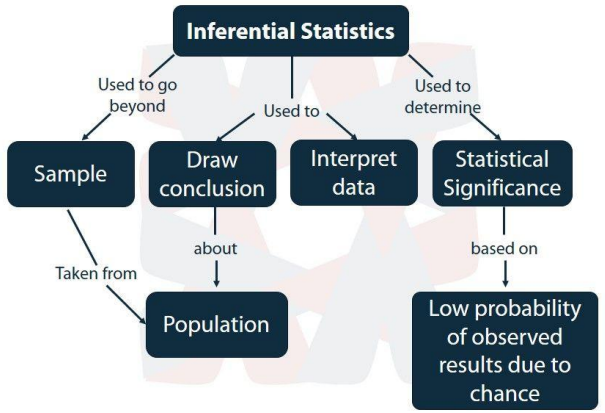
2. Binary

- Symmetric binary
- Asymmetric binary

3. Ordinal

- First, Second ,Third

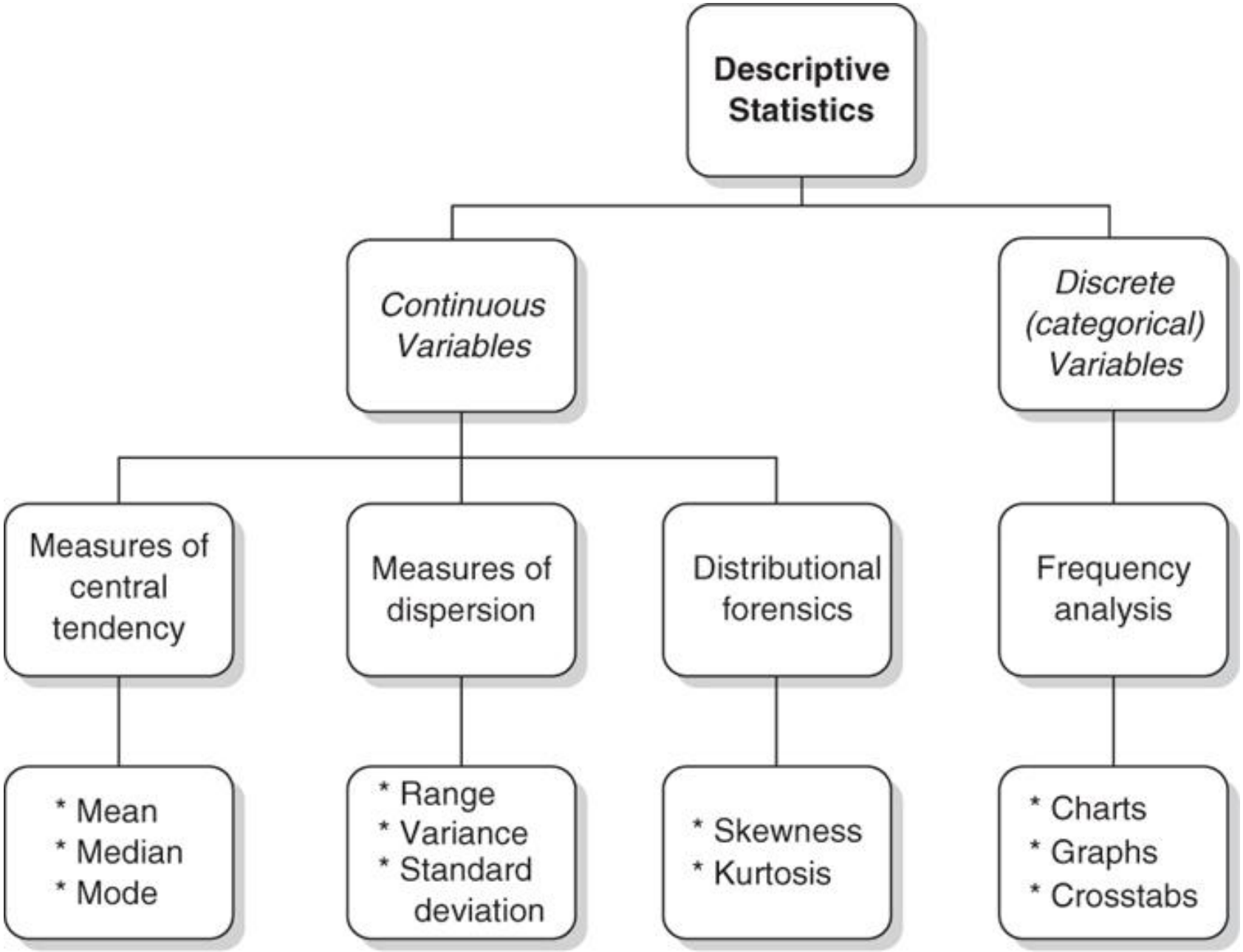




Types of statistics

- **Descriptive statistics** is concerned with Data summarization, Graphs/Charts, and tables

- **Inferential statistics** is a method used to talk about a Population parameter from a sample


Descriptive Statistics
Continuous Variables
Discrete (categorical) Variables
Measures of central tendency
Measures of dispersion
Distributional forensics
Frequency analysis
* Mean
* Median
* Mode
* Range
* Variance
* Standard deviation
* Skewness
* Kurtosis
* Charts
* Graphs
* Crosstabs

Central Tendency

Measures of Central Tendency

most *representative or typical* of all values in a group

"average"

MODE

- **most frequent data point**
- mode exists as a data point
- unaffected by extreme values
- useful for qualitative data
- may have more than 1 value

MEDIAN

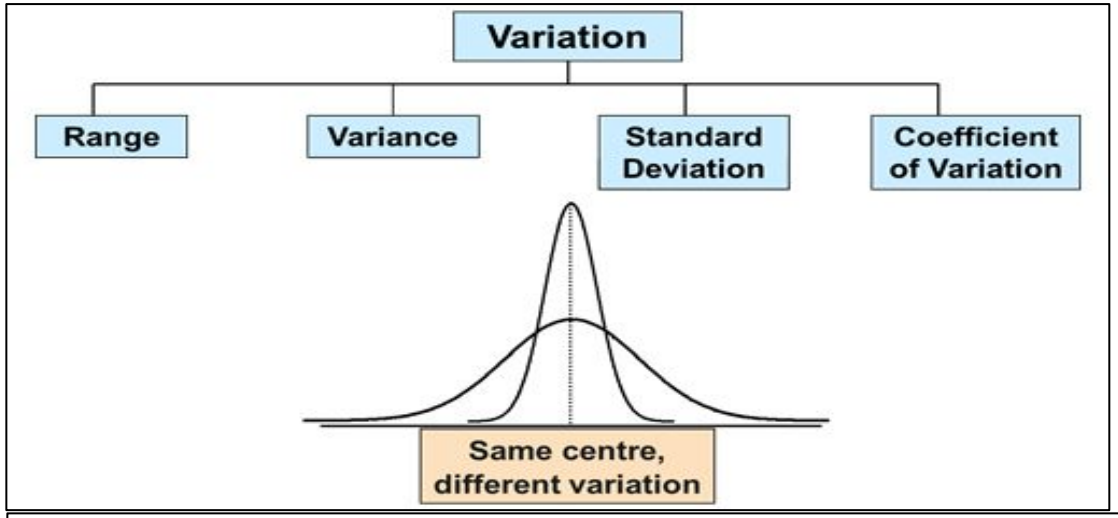
- **value that divides ranked data points into halves: 50% larger than it, 50% smaller**
- may not exist as a data point in the set
- influenced by position of items, but not their values

MEAN

$$\bar{x} = \frac{\Sigma x}{N}$$

- most stable measure
- affected by extreme values
- may not exist as a data point in the set

Measure of variation



Range

Range is the simplest of all measures of dispersion. It is calculated as the difference between maximum and minimum value in dataset.

range = X(maximum) - X (minimum)

IQR

Interquartile Range, which is a statistical measure used to describe the spread or dispersion of a data set.

- Order your data: First, arrange your data points in ascending order (from smallest to largest).
- Find the First Quartile (Q1): This is the median (middle value) of the lower half of the data. In other words, it's the median of the data points below the overall median.
- Find the Third Quartile (Q3): This is the median of the upper half of the data. It's the median of the data points above the overall median.
- Calculate the IQR: The IQR is found by subtracting the first quartile (Q1) from the third quartile (Q3). In mathematical terms,

$$IQR = Q3 - Q1.$$

- **Standard deviation**, is a measure of average spread i.e., on an average what is the difference between any data point and the central value of the variable.

$$\sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

- **Coefficient variation** is defined as ratio of standard deviation to mean.

  In symbolic form,

$$CV = \frac{S}{\bar{X}} \text{ for the sample data and } = \frac{\sigma}{\mu} \text{ for the population}$$

Example

{1, 3, 8, 3, 7, 11, 8, 3, 9, 10}

Data	Sample Mean x	Deviation $(x-\bar{x})$	Deviation ^2 $(x-\bar{x})^2$
1	6.3	-5.3	28.09
3	6.3	-3.3	10.89
8	6.3	1.7	2.89
3	6.3	-3.3	10.89
7	6.3	0.7	0.49
11	6.3	4.7	22.09
8	6.3	1.7	2.89
3	6.3	-3.3	10.89
9	6.3	2.7	7.29
10	6.3	3.7	13.69
			110.1

$= \sum (x-\bar{x})^2$

"n-1" = 9 (for denominator of sample st. deviation and variance)

Standard Deviation Calculation

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{110.1}{9}} = 3.5$$

Variance Calculation (equals standard deviation squared)

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = 12.2\overline{3}$$

Measures of Central Tendency

mean	6.3
median	7.5
mode	3

The Five Number Summary

The five numbers that help describe the center, spread and shape of data are:

- $x_{smallest}$
- First Quartile (Q_1)
- Median (Q_2)
- Third Quartile (Q_3)
- $x_{largest}$

Graphic Displays of Basic Statistical Descriptions

Boxplot: graphic display of five-number summary

Histogram: x-axis are values, y-axis repres. frequencies

Quantile plot: each value x_i is paired with f_i indicating that approximately $100\, f_i\%$ of data are $\leq x_i$

Quantile-quantile (q-q) plot: graphs the quantiles of one univariant distribution against the corresponding quantiles of another

Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane

Statistical Features- BOXPLOT

•The line in the middle is the *median* value of the data. Median is used over the mean since it is more robust to outlier values. The *first quartile* is essentially the 25th percentile; i.e 25% of the points in the data fall below that value. The *third quartile* is the 75th percentile; i.e 75% of the points in the data fall below that value. The min and max values represent the upper and lower ends of our data range.

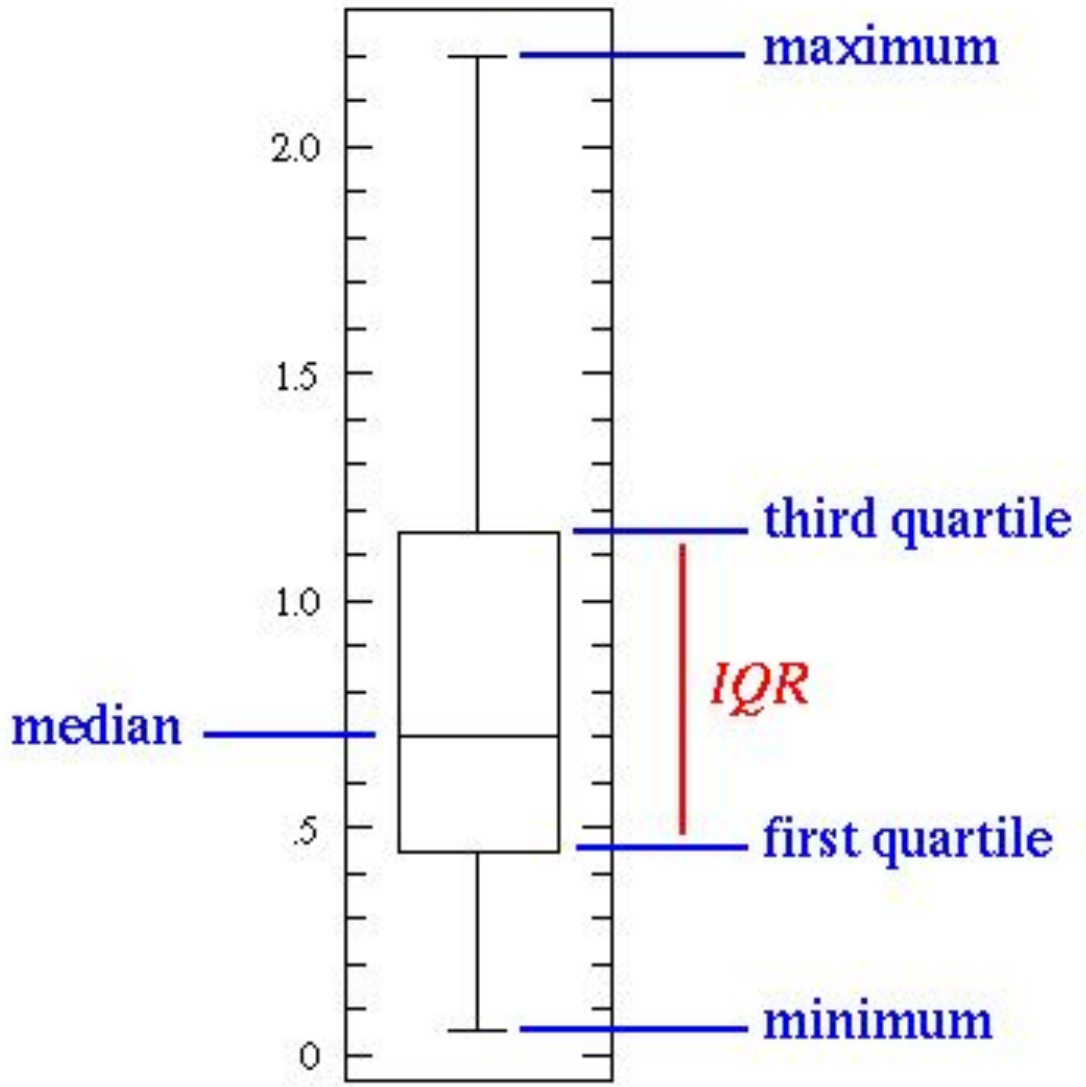
•A box plot perfectly illustrates what we can do with basic statistical features:

•When the box plot is **short** it implies that much of your data points are similar, since there are many values in a small range

•When the box plot is **tall** it implies that much of your data points are quite different, since the values are spread over a wide range

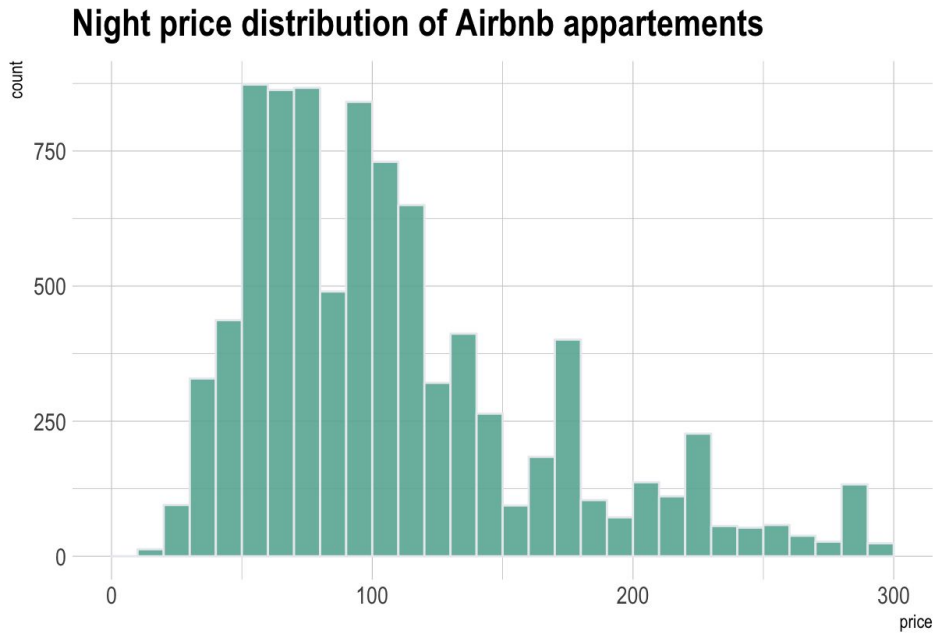
•If the median value is closer to the **bottom** then we know that most of the data has lower values. If the median value is closer to the **top** then we know that most of the data has higher values. Basically, if the median line is not in the middle of the box then it is an indication of **skewed** data.

•Are the whiskers **very long**? That means your data has a high **standard deviation** and **variance** i.e the values are spread out and highly varying. If you have long whiskers on one side of the box but not the other, then your data may be highly varying only in one direction.

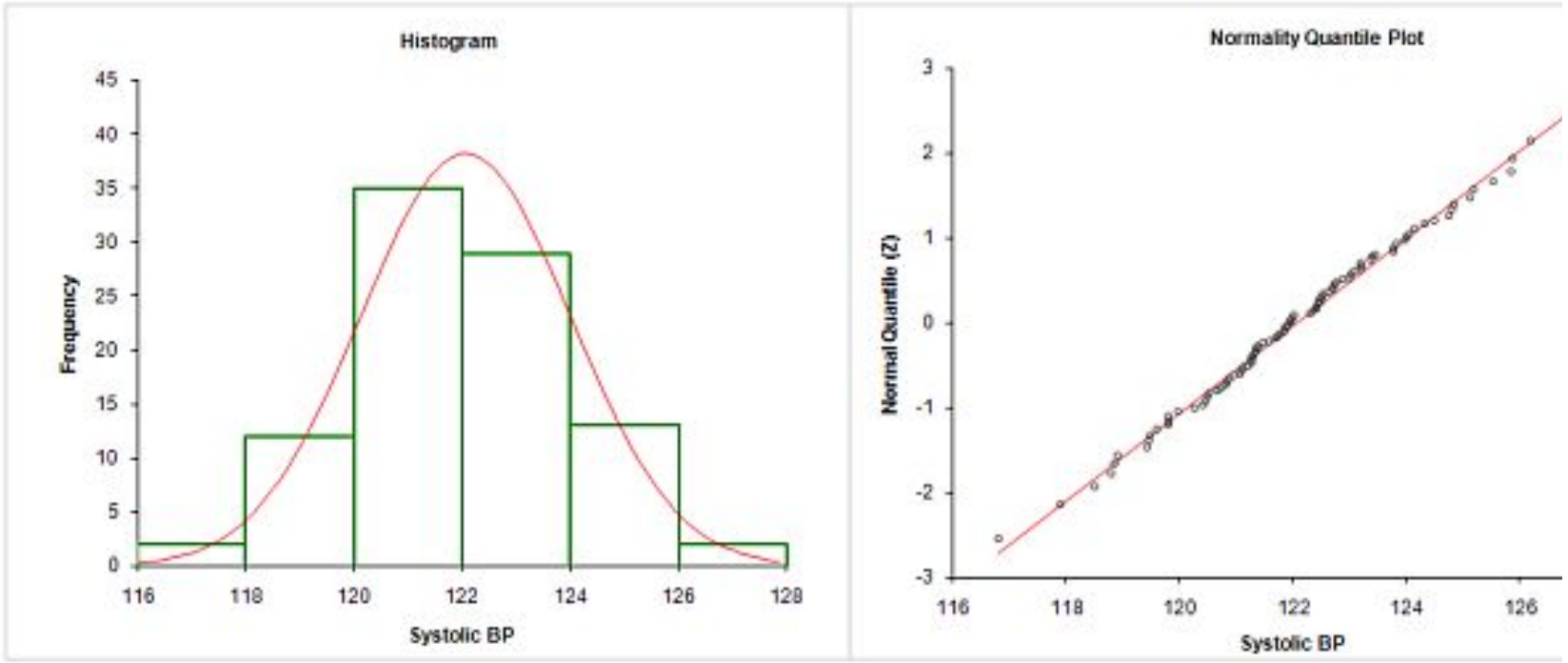


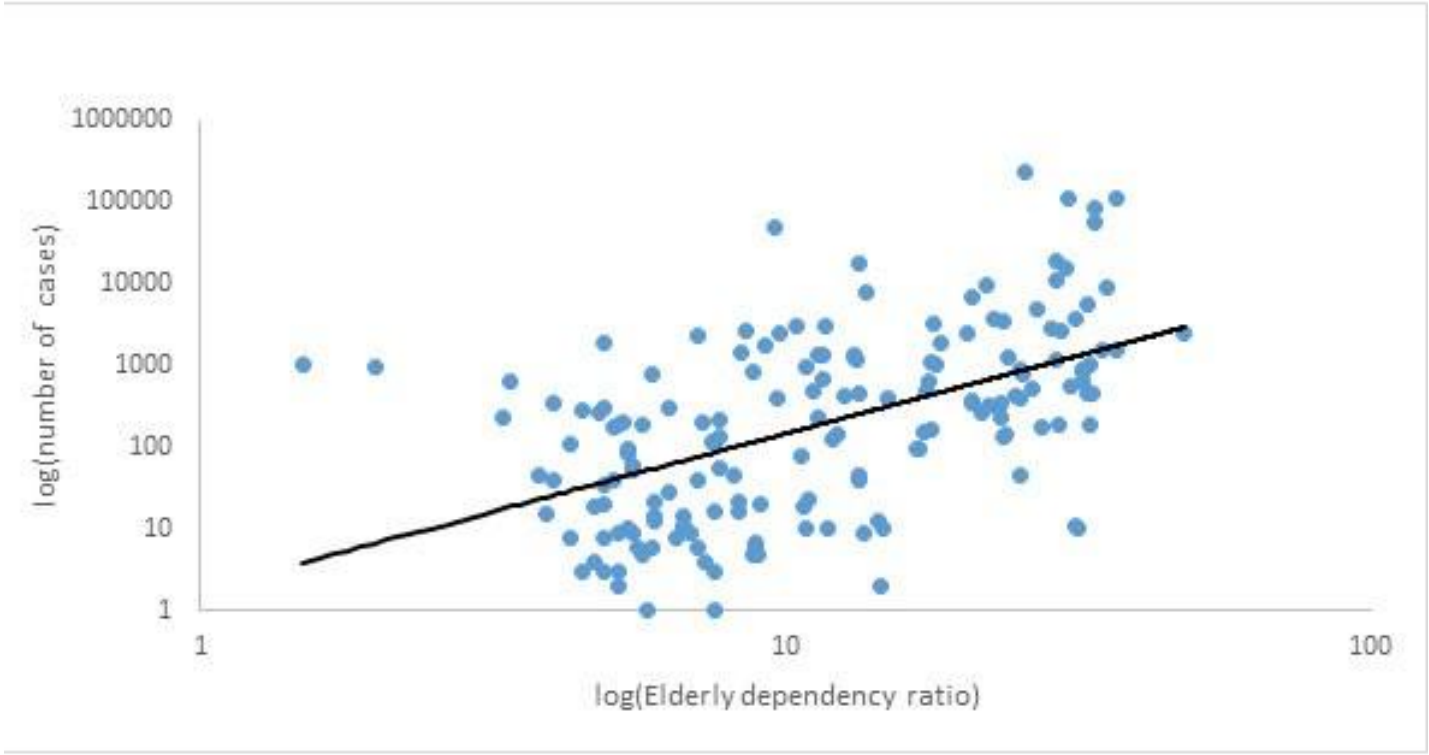
HISTOGRAM

1. Histogram is a snapshot of the frequency distribution.
2. histogram depicts the pattern of the distribution emerging from the characteristic being measured.



Quantile plot





Scatter Plot

END

Probability

Probability of an event A is defined as the ratio of two numbers m and n. In symbols

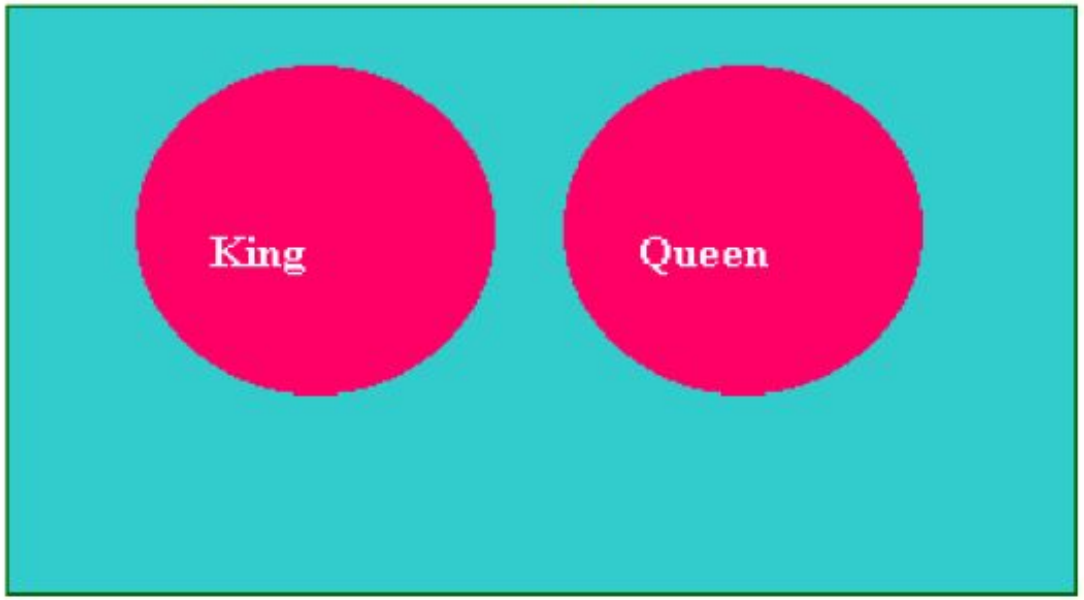
$$P(A) = \frac{m}{n}$$

where m= number of ways that are favorable to the occurrence of A and n= the total number of outcomes of the experiment (all possible outcomes)

Please note that P (A) is always > = 0 and always < = 1.
P (A) is a pure number.

Mutually Exclusive Events

Two events A and B are said to be mutually exclusive if the occurrence of A precludes the occurrence of B. For example, from a well shuffled pack of cards, if you pick up one card at random and would like to know whether it is a King or a Queen. The selected card will be either a King or a Queen. It cannot be both a King and a Queen. If King occurs, Queen will not occur and Queen occurs, King will not occur.



Independent Events

- Two events A and B are said to be independent if the occurrence of A is in no way influenced by the occurrence of B. Likewise occurrence of B is in no way influenced by the occurrence of A.

Rules for Computing Probability

1) Addition Rule -Mutually Exclusive Events

$$P(A \bigcup B) = P(A) + P(B)$$

This rule says that the probability of the union of A and B is determined by adding the probability of the events A and B.

Here the symbol $A \bigcup B$ is called A union B meaning A occurs, or B occurs or both A and B simultaneously occur. When A and B are mutually exclusive, A and B cannot simultaneously occur.

Rules for Computing Probability

2) Addition Rule –Events are not Mutually Exclusive

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

This rule says that the probability of the union of A and B is determined by adding the probability of the events A and B and then subtracting the probability of the intersection of the events A and B.

The symbol $A \cap B$ is called A intersection B meaning

both A and B simultaneously occur.

Multiplication Rule

Independent Events

$$P(A \cap B) = P(A).P(B)$$

This rule says when the two events A and B are independent, the probability of the simultaneous occurrence of A and B (also known as probability of intersection of A and B) equals the product of the probability of A and the probability of B. Of course this rule can be extended to more than two events.

Multiplication Rule

Independent Events-Example

Example:

The probability that you will get an A grade in Quantitative Methods is 0.7. The probability that you will get an A grade in Marketing is 0.5. Assuming these two courses are independent, compute the probability that you will get an A grade in both these subjects.

Solution:

Let A = getting A grade in Quantitative Methods

Let B =getting A grade in Marketing

It is given that A and B are independent.

$$P(A \cap B) = P(A).P(B) \quad = \quad 0.7.0.5 = 0.35.$$

Marginal /Join Probability - Example

A survey involving 200 families was conducted. Information regarding family income per year and whether the family buys a car are given in the following table.

Family	Income below Rs 10 Lakhs	Income of Rs. >=10 lakhs	Total
Buyer of Car	38	42	80
Non-Buyer	82	38	120
Total	120	80	200

a) What is the probability that a randomly selected family is a buyer of the car?
b) What is the probability that a randomly selected family is both a buyer of car and belonging to income of Rs. 10 lakhs and above?
c) A family selected at random is found to be belonging to income of Rs 10 lakhs and above. What is the probability that this family is buyer of car?

Solution

a) What is the probability that a randomly selected family is a buyer of the Car?

- 80/200 =0.40.

b) What is the probability that a randomly selected family is both a buyer of car and belonging to income of Rs. 10 lakhs and above?

- 42/200 =0.21.

c) A family selected at random is found to be belonging to income of Rs 10 lakhs and above. What is the probability that this family is buyer of car?

- 42/80 =0.525. Note this is a case of conditional probability of buyer given income is Rs. 10 lakhs and above.

Probability

- **Joint probability**
- It is the probability of multiple events occurring together. For eg:
  - Probability of drawing a king from a deck of cards is 4/52
  - Probability of drawing a red colour card from a deck of cards is 26/52
  - Probability of drawing a red colour king = 2/52

- **Conditional probability**
- It is the probability that an event has occurred given another event has occurred. For eg:
  - Given the card drawn is red (an event has occurred)
  - What is the probability it is a king (event not yet observed)?
  - Since the card is red, there are 26 likely values for red.
  - Of these 26 possible values we are interested in king which is 2 (king of diamonds and heart)

Bayesian Statistics

Likelihood of the evidence 'E' if the Hypothesis 'H' is true

Prior Probability

$$P(H|E) = \frac{P(H) * P(E|H)}{P(E)}$$

Posterior Probability of 'H' given the evidence

Priori probability that the evidence itself is true

Baye's Theorem

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_k)P(B_k)}$$

- where:

  B_i = i^{th} event of k mutually exclusive and collectively exhaustive events

  A = new event that might impact $P(B_i)$

Examples

Bayesian filtering allows us to predict the chance a message is really spam given the "test results" (the presence of certain words). Clearly, words like "viagra" have a higher chance of appearing in spam messages than in normal ones.

Spam filtering based on a blacklist is flawed — it's too restrictive and false positives are too great. But Bayesian filtering gives us a middle ground — we use *probabilities*. As we analyze the words in a message, we can compute the chance it is spam (rather than making a yes/no decision). If a message has a 99.9% chance of being spam, it probably is. As the filter gets trained with more and more messages, it updates the probabilities that certain words lead to spam messages. Advanced Bayesian filters can examine multiple words in a row, as another data point.

Probability and Statistics

Probability is the chance of an outcome in an experiment (also called event).

Event: Tossing a fair coin
Outcome: Head, Tail

Probability deals with **predicting** the likelihood of **future** events.	Statistics involves the **analysis** of the **frequency** of **past** events

Example: Consider there is a drawer containing 100 socks: 30 red, 20 blue and 50 black socks.

We can use probability to answer questions about the selection of a random sample of these socks.

- **PQ1.** What is the probability that we draw two blue socks or two red socks from the drawer?
- **PQ2.** What is the probability that we pull out three socks or have matching pair?
- **PQ3.** What is the probability that we draw five socks and they are all black?

Statistics

Instead, if we have no knowledge about the type of socks in the drawers, then we enter into the realm of statistics. Statistics helps us to infer properties about the population on the basis of the random sample.

Questions that would be statistical in nature are:

- **SQ1**: A random sample of 10 socks from the drawer produced one blue, four red, five black socks. What is the total population of black, blue or red socks in the drawer?
- **SQ2**: We randomly sample 10 socks, and write down the number of black socks and then return the socks to the drawer. The process is done for five times. The mean number of socks for each of these trial is 7. What is the true number of black socks in the drawer?
- etc.

Probability vs. Statistics

In other words:

- In probability, we are given a model and asked what kind of data we are likely to see.
- In statistics, we are given data and asked what kind of model is likely to have generated it.

Example 4.1: Measles Study

- A study on health is concerned with the incidence of childhood measles in parents of childbearing age in a city. For each couple, we would like to know how likely, it is that either the mother or father or both have had childhood measles.
- The current census data indicates that 20% adults between the ages 17 and 35 (regardless of sex) have had childhood measles.
    - This give us the probability that an individual in the city has had childhood measles.

Defining Random Variable

Definition 4.1: Random Variable

A random variable is a rule that assigns a numerical value to an outcome of interest.

Example 4.2: In "measles Study", we define a random variable X as the number of parents in a married couple who have had childhood measles.

This random variable can take values of $0, 1$ *and* 2.

Note:

- Random variable is not exactly the same as the variable defining a data.
- The probability that the random variable takes a given value can be computed using the rules governing probability.
  - For example, the probability that $X = 1$ means either mother or father but not both has had measles is 0.32. Symbolically, it is denoted as $\mathbf{P(X=1) = 0.32}$

Probability Distribution

Definition 4.2: Probability distribution

A probability distribution is a definition of probabilities of the values of random variable.

-

Example 4.3: Given that 0.2 is the probability that a person (in the ages between 17 and 35) has had childhood measles. Then the probability distribution is given by

X	Probability
0	0.64
1	0.32
2	0.04

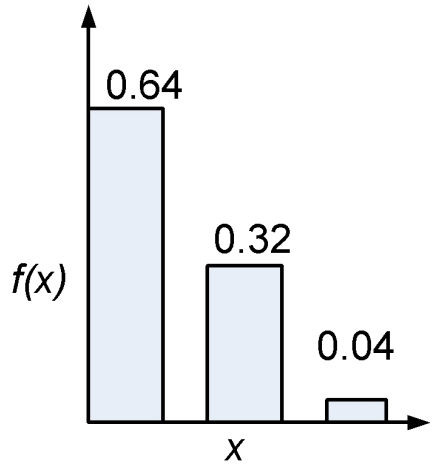
?

Probability Distribution

- In data analytics, the probability distribution is important with which many statistics making inferences about population can be derived .
  - In general, a probability distribution function takes the following form

Example: Measles Study

	0	1	2
	0.64	0.32	0.04



Taxonomy of Probability Distributions

Discrete probability distributions

- Binomial distribution
- Multinomial distribution
- Poisson distribution
- Hypergeometric distribution

Continuous probability distributions

- Normal distribution
- Standard normal distribution
- Gamma distribution
- Exponential distribution
- Chi square distribution
- Lognormal distribution
- Weibull distribution

Usage of Probability Distribution

- Distribution (discrete/continuous) function is widely used in simulation studies.
  - A simulation study uses a computer to simulate a real phenomenon or process as closely as possible.
  - The use of simulation studies can often eliminate the need of costly experiments and is also often used to study problems where actual experimentation is impossible.

Examples 4.4:

1) A study involving testing the effectiveness of a new drug, the number of cured patients among all the patients who use such a drug approximately follows a binomial distribution.

2) Operation of ticketing system in a busy public establishment (e.g., airport), the arrival of passengers can be simulated using Poisson distribution.

Discrete Probability Distributions

Binomial Distribution

- In many situations, an outcome has only two outcomes: success and failure.
  - Such outcome is called dichotomous outcome.
- An experiment when consists of repeated trials, each with dichotomous outcome is called Bernoulli process. Each trial in it is called a Bernoulli trial.

Example 4.5: Firing bullets to hit a target.

- Suppose, in a Bernoulli process, we define a random variable $X \equiv$ the number of successes in trials.
- Such a random variable obeys the binomial probability distribution, if the experiment satisfies the following conditions:

  1) The experiment consists of n trials.

  2) Each trial results in one of two mutually exclusive outcomes, one labelled a *"success"* and the other a *"failure"*.

  3) The probability of a success on a single trial is equal to $\boldsymbol{p}$. The value of p remains constant throughout the experiment.

  4) The trials are independent.

Defining Binomial Distribution

Definition 4.3: **Binomial distribution**

The function for computing the probability for the binomial probability distribution is given by

$$f(x) = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$$

for x = 0, 1, 2, …., n

Here, $f(x) = P(X = x)$, where X denotes "the number of success" and $X = x$ denotes the number of success in x trials.

Binomial Distribution

Example 4.6: Measles study

X = having had childhood measles a success

$p = 0.2$, the probability that a parent had childhood measles

$n = 2$, here a couple is an experiment and an individual a trial, and the number of trials is two.

Thus,

$$P(x = 0) = \frac{2!}{0!(2-0)!}(0.2)^0(0.8)^{2-0} = \mathbf{0.64}$$

$$P(x = 1) = \frac{2!}{1!\,(2-1)!}(0.2)^1(0.8)^{2-1} = \mathbf{0.32}$$

$$P(x = 2) = \frac{2!}{2!\,(2-2)!}(0.2)^2(0.8)^{2-2} = \mathbf{0.04}$$

Binomial Distribution

Example 4.7: Verify with real-life experiment

Suppose, 10 pairs of random numbers are generated by a computer (Monte-Carlo method)

15 38 68 39 49 54 19 79 38 14

If the value of the digit is 0 or 1, the outcome is "had childhood measles", otherwise, (digits 2 to 9), the outcome is "did not".

For example, in the first pair (i.e., 15), representing a couple and for this couple, $x = 1$. The frequency distribution, for this sample is

x	0	1	2
$f(x)=P(X=x)$	0.7	0.3	0.0

Note: This has close similarity with binomial probability distribution!

The Multinomial Distribution

The binomial experiment becomes a multinomial experiment, if we let each trial has more than two possible outcome.

Definition 4.4: **Multinomial distribution**

If a given trial can result in the ***k*** outcomes $E_1, E_2, \dots\dots, E_k$ with probabilities $p_1, p_2, \dots\dots, p_k$, then the probability distribution of the random variables $X_1, X_2, \dots\dots, X_k$ representing the number of occurrences for $E_1, E_2, \dots\dots, E_k$ in n independent trials is

$$f(x_1, x_2, \dots\dots, x_k) = \binom{n}{x_1, x_2, \dots\dots, x_k} {p_1}^{x_1} {p_2}^{x_2} \dots\dots {p_k}^{x_k}$$

where $\binom{n}{x_1, x_2, \dots\dots, x_k} = \frac{n!}{x_1! x_2! \dots\dots x_k!}$

$$\sum_{i=1}^{k} x_i = n \text{ and } \sum_{i=1}^{k} p_i = 1$$

The Hypergeometric Distribution

- Collection of samples with two strategies
  - With replacement
  - Without replacement
- A necessary condition of the binomial distribution is that all trials are **independent** to each other.
  - When sample is collected "with replacement", then each trial in sample collection is independent.

Example 4.8:
Probability of observing three red cards in 5 draws from an ordinary deck of 52 playing cards.
- You draw one card, note the result and then returned to the deck of cards
- Reshuffled the deck well before the next drawing is made

- The hypergeometric distribution *does not require* independence and is based on the sampling done **without** replacement.

The Hypergeometric Distribution

- In general, the hypergeometric probability distribution enables us to find the probability of selecting x successes in n trials from N items.

Properties of Hypergeometric Distribution

- A random sample of size n is selected without replacement from N items.
- k of the N items may be classified as success and $N-k$ items are classified as failure.

Let X denotes a hypergeometric random variable defining the number of successes.

Definition 4.5: Hypergeometric Probability Distribution

The probability distribution of the hypergeometric random variable X, the number of successes in a random sample of size n selected from N items of which k are labelled success and $N-k$ labelled as failure is given by

$$f(x) = P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

$$\max(0, n-(N-k)) \leq x \leq \min(n, k)$$

Multivariate Hypergeometric Distribution

The hypergeometric distribution can be extended to treat the case where the N items can be divided into k classes $A_1, A_2, \ldots\ldots, A_k$ with a_1 elements in the first class A_1, … and a_k elements in the k^{th} class. We are now interested in the probability that a random sample of size n yields x_1 elements from A_1, x_2 elements from $A_2, \ldots\ldots, x_k$ elements from A_k.

Definition 4.6: Multivariate Hypergeometric Distribution

If N items are partitioned into k classes $a_1, a_2, \ldots\ldots, a_k$ respectively, then the probability distribution of the random variables $X_1, X_2, \ldots\ldots, X_k$, representing the number of elements selected from $A_1, A_2, \ldots\ldots, A_k$ in a random sample of size n, is

$$f(x_1, x_2, \ldots\ldots, x_k) = P(X_1 = x_1, X_2 = x_2, \ldots\ldots, X_k = x_k) = \frac{\binom{a_1}{x_1}\binom{a_2}{x_2}\ldots\ldots\binom{a_k}{x_k}}{\binom{N}{n}}$$

$$\text{with } \sum_{i=1}^{k} x_i = n \text{ and } \sum_{i=1}^{k} a_i = N$$

The Poisson Distribution

There are some experiments, which involve the occurring of the number of outcomes during a given time interval (or in a region of space).

Such a process is called Poisson process.

Example 4.9:

Number of clients visiting a ticket selling counter in a metro station.

The Poisson Distribution

Properties of Poisson process

- The number of outcomes in one time interval is independent of the number that occurs in any other disjoint interval [Poisson process has no memory]
- The probability that a single outcome will occur during a very short interval is proportional to the length of the time interval and does not depend on the number of outcomes occurring outside this time interval.
- The probability that more than one outcome will occur in such a short time interval is negligible.

Definition 4.7: **Poisson distribution**

The probability distribution of the Poisson random variable X, representing the number of outcomes occurring in a given time interval t, is

$$f(x, \lambda t) = P(X = x) = \frac{e^{-\lambda t}.(\lambda t)^x}{x!}, x = 0, 1, \dots \dots$$

where λ is the average number of outcomes per unit time and $e = 2.71828 \dots$

Descriptive measures

Given a random variable X in an experiment, we have denoted $f(x) = P(X = x)$, the probability that $X = x$. For discrete events $f(x) = 0$ for all values of x except $x = 0, 1, 2, \ldots\ldots$

Properties of discrete probability distribution

1. $0 \leq f(x) \leq 1$
2. $\sum f(x) = 1$
3. $\mu = \sum x.f(x)$ [is the mean]
4. $\sigma^2 = \sum(x - \mu)^2 . f(x)$ [is the variance]

In 2, 3 *and* 4, summation is extended for all possible discrete values of x.

Note: For discrete uniform distribution, $f(x) = \frac{1}{n}$ with $x = 1, 2, \ldots\ldots, n$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

Descriptive measures

1. **Binomial distribution**

The binomial probability distribution is characterized with p (the probability of success) and n (is the number of trials). Then

$$\mu = n.p$$

$$\sigma^2 = np(1-p)$$

2. **Hypergeometric distribution**

The hypergeometric distribution function is characterized with the size of a sample (n), the number of items (N) and k labelled success. Then

$$\mu = \frac{nk}{N}$$

$$\sigma^2 = \frac{N-n}{N-1}.\mathrm{n}.\frac{k}{N}(1-\frac{k}{N})$$

Descriptive measures

3. **Poisson Distribution**

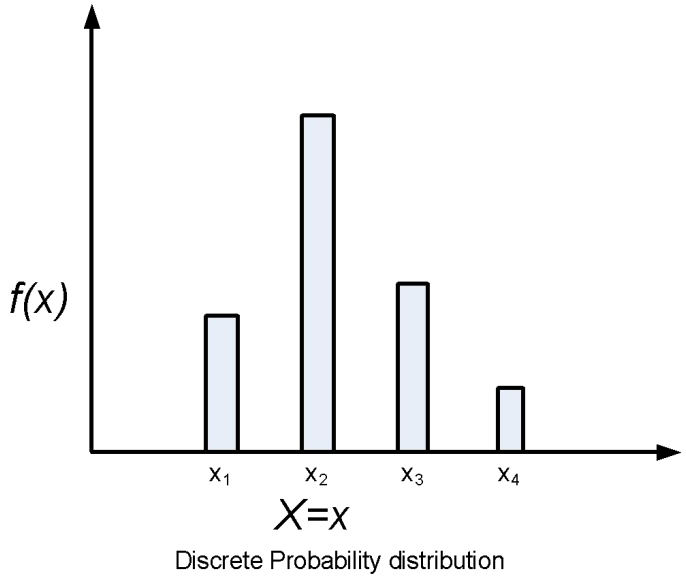
   The Poisson distribution is characterized with λt where $\lambda = the\ mean\ of\ outcomes$ and $t = time\ interval$.

$$\mu = \lambda t$$

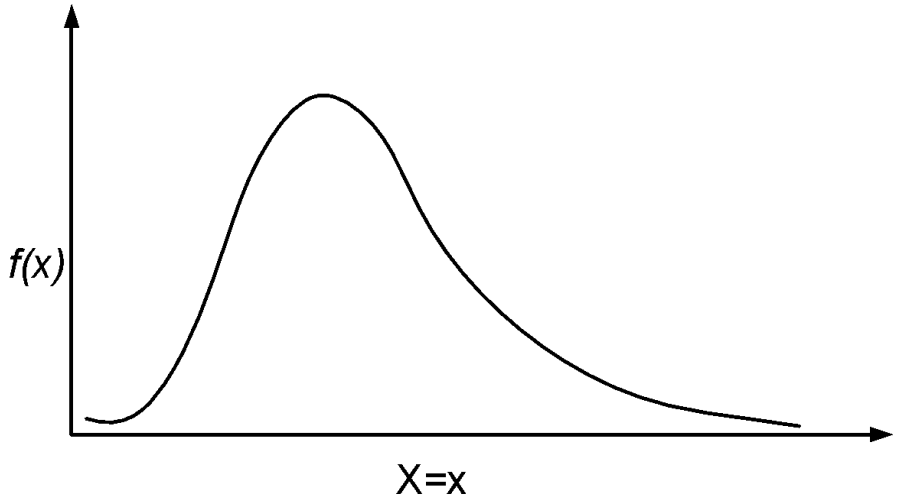
$$\sigma^2 = \lambda t$$

Continuous Probability Distributions

Continuous Probability Distributions



Discrete Probability distribution



Continuous Probability Distribution

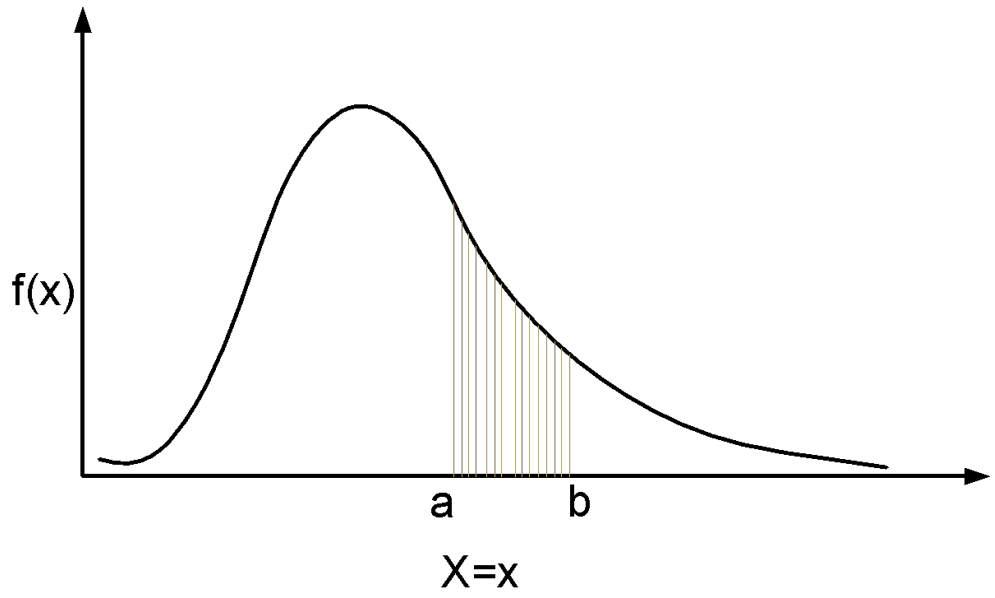
Continuous Probability Distributions

- When the random variable of interest can take any value in an interval, it is called continuous random variable.
    - Every continuous random variable has an infinite, uncountable number of possible values (i.e., any value in an interval)
- Consequently, continuous random variable differs from discrete random variable.

Properties of Probability Density Function

The function $f(x)$ is a probability density function for the continuous random variable X, defined over the set of real numbers R, if

1. $f(x) \geq 0$, for all $x \in R$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$
4. $\mu = \int_{-\infty}^{\infty} xf(x)\,dx$
5. $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$



Continuous Uniform Distribution

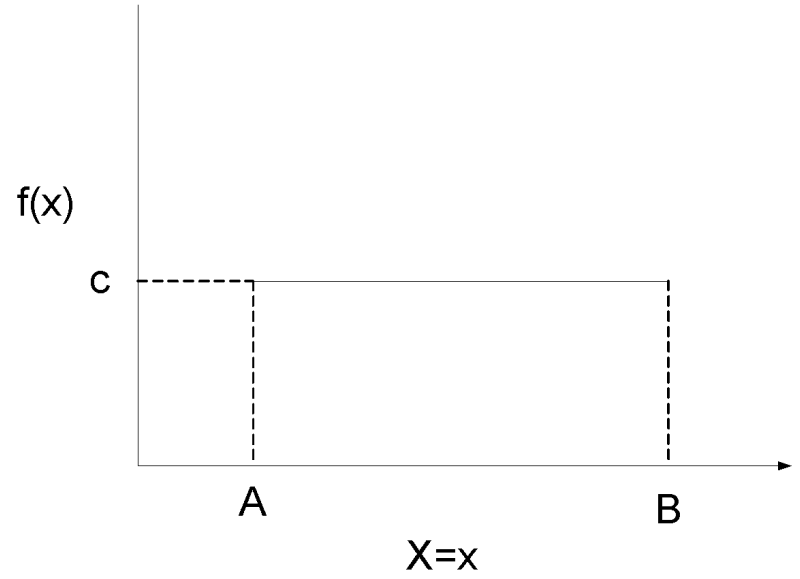
- One of the simplest continuous distribution in all of statistics is the continuous uniform distribution.

Definition 4.8: **Continuous Uniform Distribution**

The density function of the continuous uniform random variable X on the interval $[A, B]$ is:

$$f(x: A, B) = \begin{cases} \dfrac{1}{B - A} & A \leq x \leq B \\ 0 & Otherwise \end{cases}$$

Continuous Uniform Distribution



Note:

a) $\int_{\infty}^{-\infty} f(x)dx = \frac{1}{B-A} \times (B-A) = 1$

b) $P(c < x < d) = \frac{d-c}{B-A}$ where both c and d are in the interval (A,B)

c) $\mu = \frac{A+B}{2}$

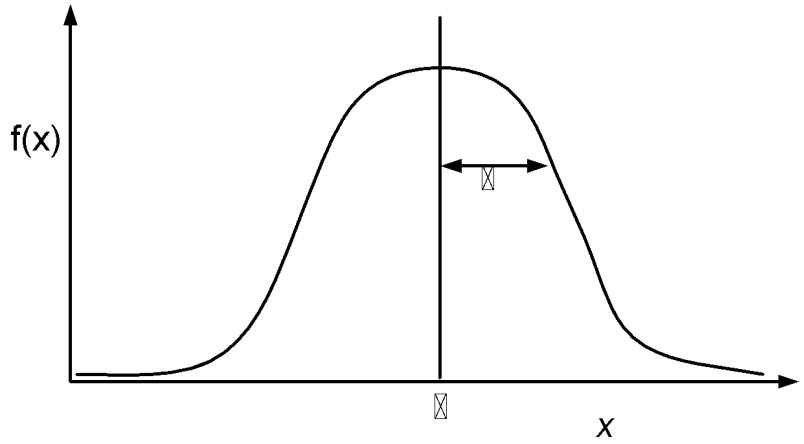
d) $\sigma^2 = \frac{(B-A)^2}{12}$

Normal Distribution

- The most often used continuous probability distribution is the normal distribution; it is also known as **Gaussian** distribution.
- Its graph called the normal curve is the bell-shaped curve.
- Such a curve approximately describes many phenomenon occur in nature, industry and research.
    - Physical measurement in areas such as meteorological experiments, rainfall studies and measurement of manufacturing parts are often more than adequately explained with normal distribution.
- A continuous random variable X having the bell-shaped distribution is called a normal random variable.

Normal Distribution

- The mathematical equation for the probability distribution of the normal variable depends upon the two parameters μ and σ, its mean and standard deviation.



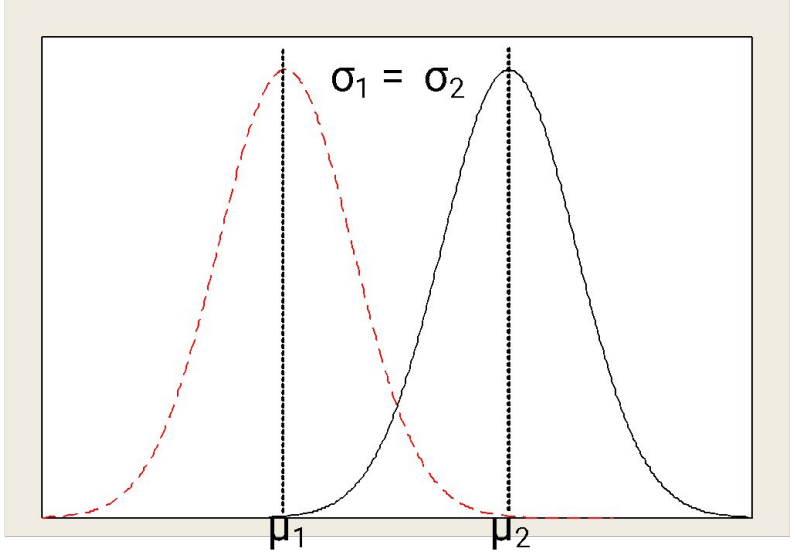
Definition 4.9: Normal distribution

The density of the normal variable x with mean μ and variance σ^2 is

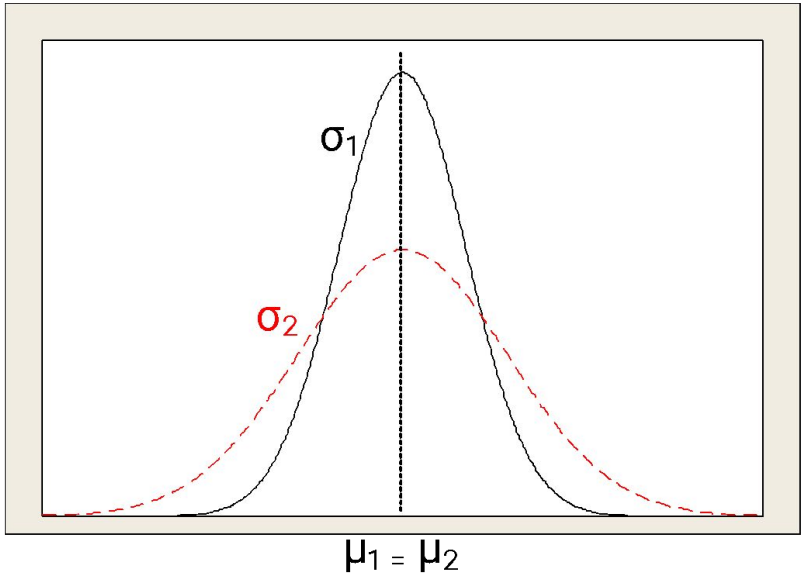
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty$$

where $\pi = 3.14159$ ... and $e = 2.71828$ ....., the Naperian constant

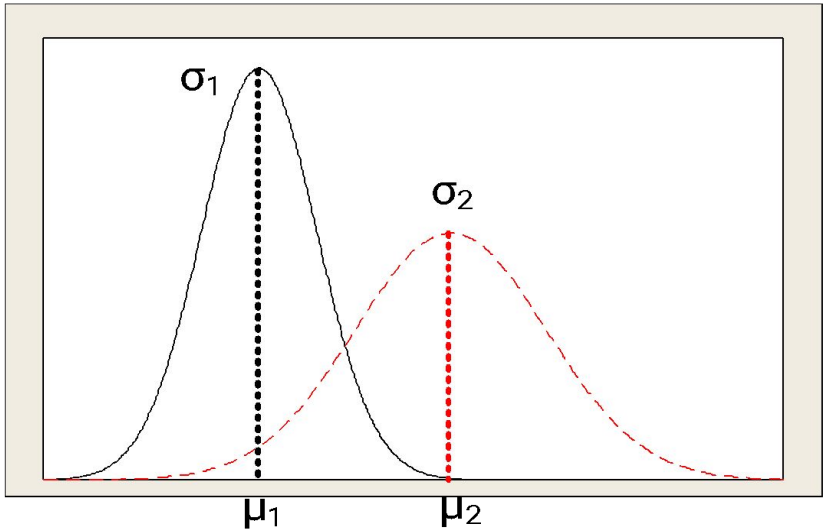
Normal Distribution



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 = \sigma_2$



Normal curves with $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$

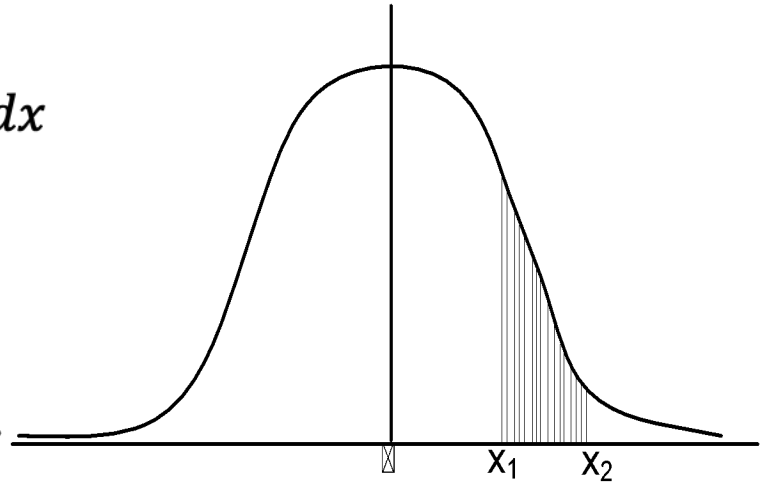
Properties of Normal Distribution

- The curve is symmetric about a vertical axis through the mean μ.
- The random variable x can take any value from $-\infty \; to \; \infty$.
- The most frequently used descriptive parameter s define the curve itself.
- The mode, which is the point on the horizontal axis where the curve is a maximum occurs at $x = \mu$.
- The total area under the curve and above the horizontal axis is equal to 1.

$$\int_{-\infty}^{\infty} f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = 1$$

- $$\mu = \int_{-\infty}^{\infty} x.f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x.e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$
- $$\sigma^2 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2.e^{-\frac{1}{2}[(x-\mu)/\sigma^2]} dx$$
- $$P(x_1 < x < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

denotes the probability of x in the interval (x_1, x_2).



Standard Normal Distribution

- The normal distribution has computational complexity to calculate $P(x_1 < x < x_2)$ for any two (x_1, x_2) and given μ and σ
- To avoid this difficulty, the concept of z-transformation is followed.

- X: Normal distribution with mean μ and variance σ^2.
- Z: Standard normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$.
- Therefore, if *f(x)* assumes a value, then the corresponding value of $f(z)$ is given by

$$f(x:\mu,\sigma): \quad P(x_1 < x < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

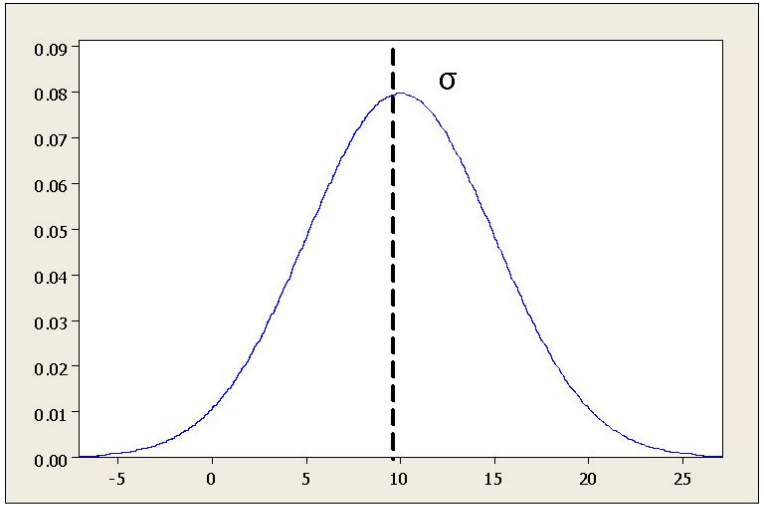
$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2} dz$$

$$= f(z:0,\sigma)$$

Standard Normal Distribution

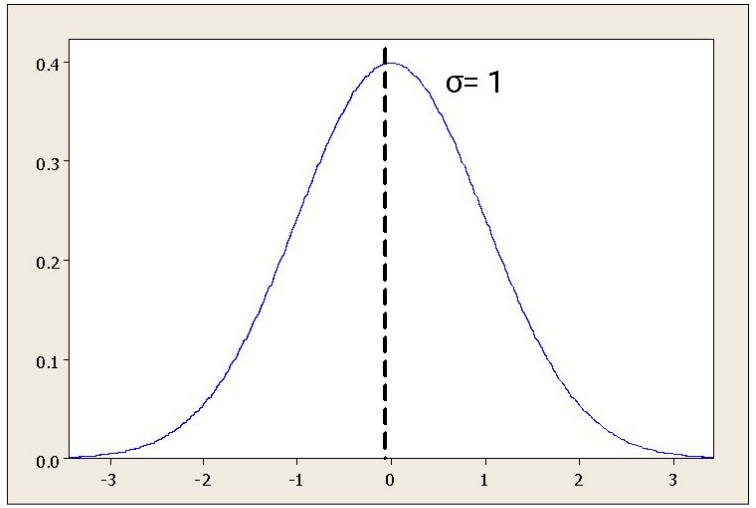
Definition 4.10: **Standard normal distribution**

The distribution of a normal random variable with mean 0 and variance 1 is called a standard normal distribution.



$x=\mu$

$f(x: \mu, \sigma)$



$\mu=0$

$f(z: 0, 1)$

Gamma Distribution

The gamma distribution derives its name from the well known gamma function in mathematics.

Definition 4.11: Gamma Function

$$\Gamma(\alpha) = \int_0^{\alpha} x^{\alpha-1} e^{-x} dx \quad \text{for } \alpha > 0$$

Integrating by parts, we can write,

$$\Gamma(\alpha) = (\alpha - 1) \int_0^{\alpha} x^{\alpha-2} e^{-x} dx$$
$$= (\alpha - 1)\Gamma(\alpha - 1)$$

Thus Γ function is defined as a recursive function.

Gamma Distribution

When $\alpha = n$, we can write,

$$\Gamma(n) = (n-1)(n-2) \dots \dots \dots . \Gamma(1)$$

$$= (n-1)(n-2) \dots \dots \dots .3.2.1$$

$$= (n-1)!$$

Further, $\Gamma(1) = \int_0^\infty e^{-x} dx = 1$

Note:

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$ [An important property]

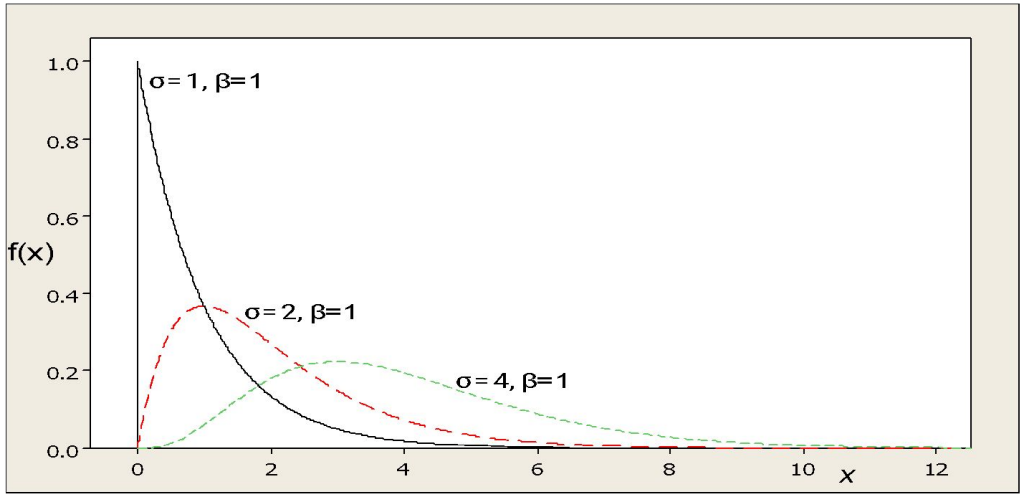
Gamma Distribution

Definition 4.12: **Gamma Distribution**

The continuous random variable x has a gamma distribution with parameters α and β such that:

$$f(x:\alpha,\beta) = \begin{cases} \frac{1}{\beta^{\alpha}\,\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} & x > 0 \\ 0 & Otherwise \end{cases}$$

where $\alpha > 0$ and $\beta > 0$



Exponential Distribution

Definition 4.13: **Exponential Distribution**

The continuous random variable x has an exponential distribution with parameter β , where:

$$f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}} & \text{where } \beta > 0 \\ 0 \end{cases}$$

Note:

1) The mean and variance of gamma distribution are

$$\mu = \alpha\beta$$
$$\sigma^2 = \alpha\beta^2$$

2) The mean and variance of exponential distribution are

$$\mu = \beta$$
$$\sigma^2 = \beta^2$$

Chi-Squared Distribution

Definition 4.14: **Chi-squared distribution**

The continuous random variable x has a Chi-squared distribution with v degrees of freedom, is given by

$$f(x:v) = \begin{cases} \frac{1}{2^{\frac{v}{2}}\,\Gamma(v/2)} x^{v/2-1} e^{-\frac{x}{2}}, & x > 0 \\ 0 & Otherwise \end{cases}$$

where v is a positive integer.

- The Chi-squared distribution plays an important role in statistical inference .
- The mean and variance of Chi-squared distribution are:

$$\mu = v \text{ and } \sigma^2 = 2v$$

Lognormal Distribution

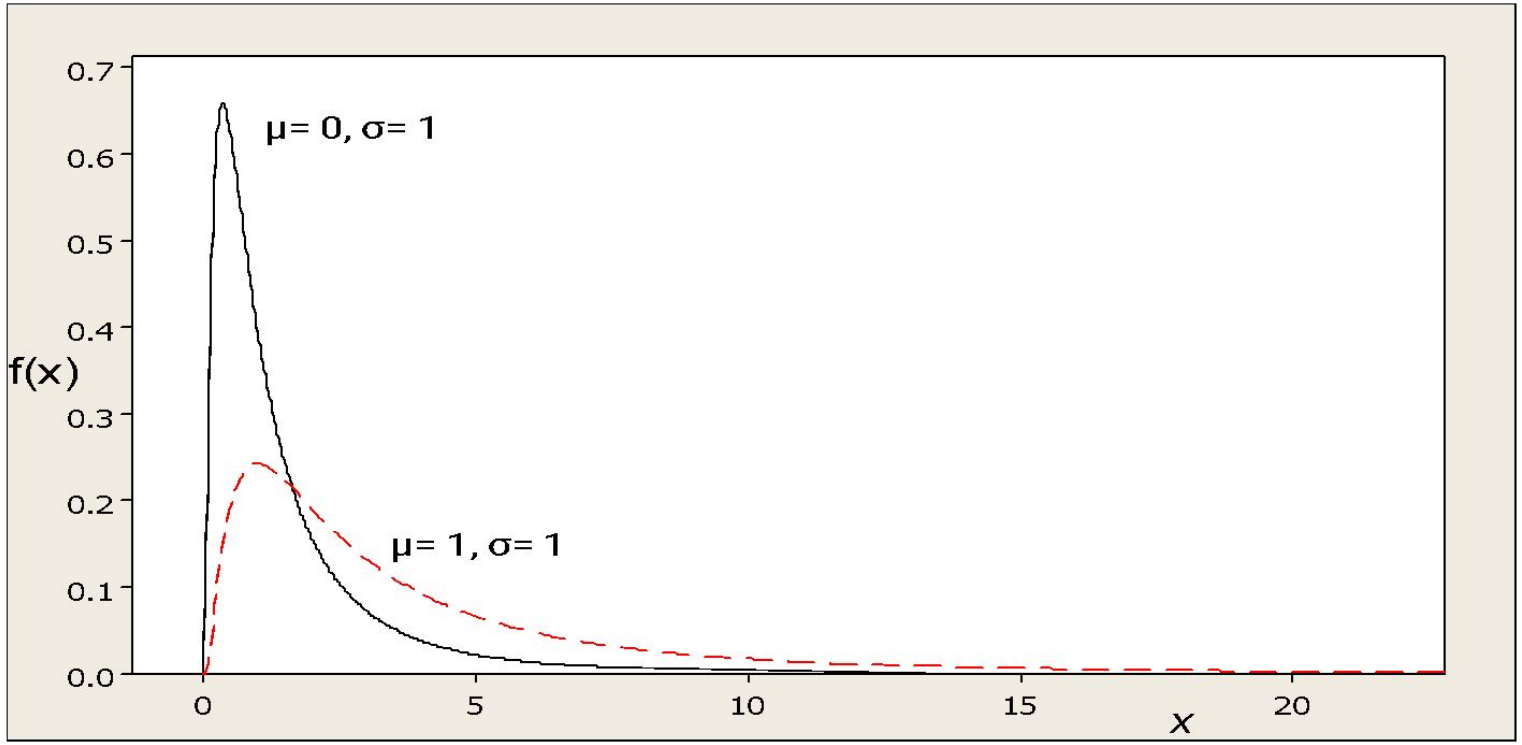
The lognormal distribution applies in cases where a natural log transformation results in a normal distribution.

Definition 4.15: **Lognormal distribution**

The continuous random variable x has a lognormal distribution if the random variable $y = \ln(x)$ has a normal distribution with mean μ and standard deviation σ. The resulting density function of x is:

$$f(x:\mu,\sigma) = \begin{cases} \frac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[\ln(x)-\mu]^2} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Lognormal Distribution



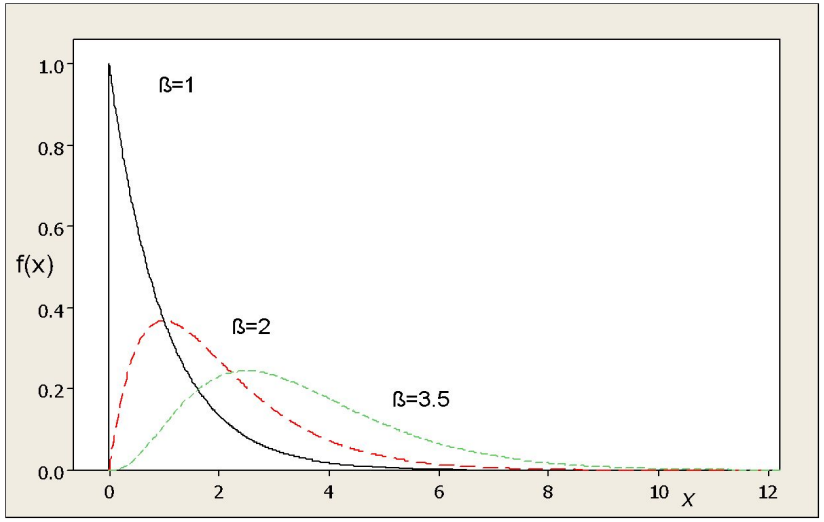
Weibull Distribution

Definition 4.16: Weibull Distribution

The continuous random variable x has a Weibull distribution with parameter α and β such that.

$$f(x:\alpha,\beta) = \begin{cases} \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}} & x > 0 \\ 0 & Otherwise \end{cases}$$

where $\alpha > 0$ and $\beta > 0$



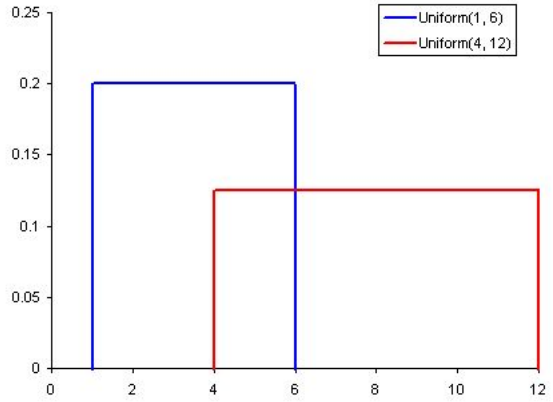
The mean and variance of Weibull distribution are:

$$\mu = \alpha^{-1/\beta}\Gamma\left(1+\frac{1}{\beta}\right)$$

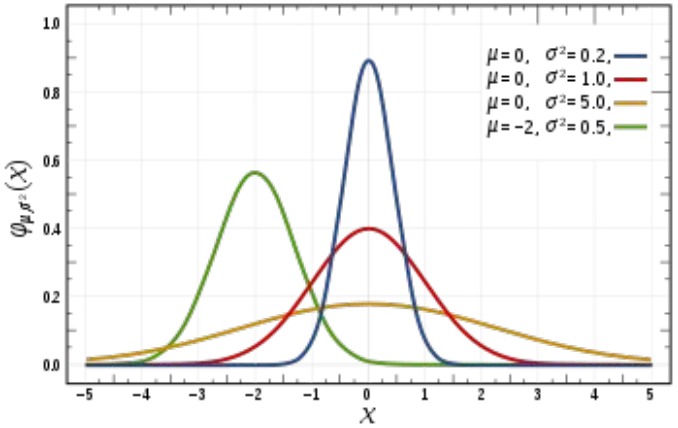
$$\sigma^2 = \alpha^{-2/\beta}\{\Gamma\left(1+\frac{2}{\beta}\right) - [\Gamma(1+\frac{1}{\beta})]^2\}$$

Probability Distributions

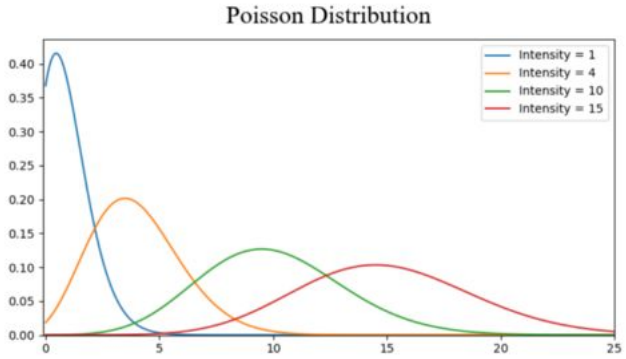
A **Uniform Distribution** is the most basic of the 3 we show here. It has a single value which only occurs in a certain range while anything outside that range is just 0. It's very much an "on or off" distribution. We can also think of it as an indication of a categorical variable with 2 categories: 0 or the value. Your categorical variable might have multiple values other than 0 but we can still visualize it in the same was as a piecewise function of multiple uniform distributions.



- A **Normal Distribution**, commonly referred to as a **Gaussian Distribution**, is specifically defined by its mean and standard deviation. The mean value shifts the distribution spatially and the standard deviation controls the spread. The import distinction from other distributions (e.g poisson) is that the standard deviation is the same in all directions. Thus with a Gaussian distribution we know the average value of our dataset as well as the spread of the data i.e is it spread over a wide range or is it highly concentrated around a few values.



- A **Poisson Distribution** is similar to the Normal but with an added factor of *skewness*. With a low value for the skewness a poisson distribution will have relatively uniform spread in all directions just like the Normal. But when the skewness value is high in magnitude then the spread of our data will be different in different directions; in one direction it will be very spread and in the other it will be highly concentrated.

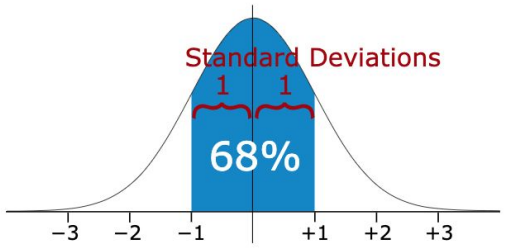


Normal Distribution

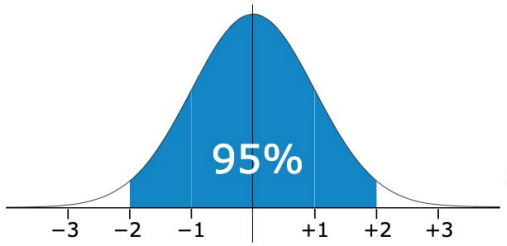
Standard Deviations

The Standard Deviation is a measure of how spread out numbers are (read that page for details on how to calculate it).

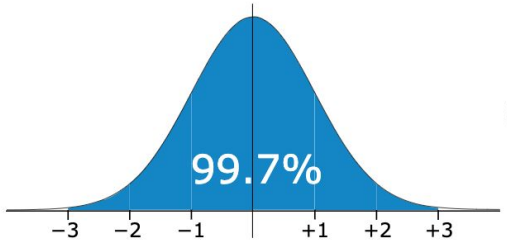
When we calculate the standard deviation we find that **generally**:



68% of values are within **1 standard deviation** of the mean



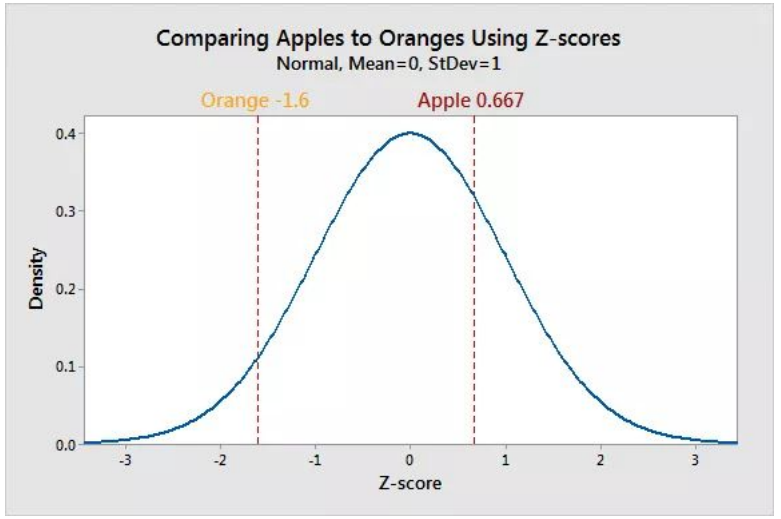
95% of values are within **2 standard deviations** of the mean



99.7% of values are within **3 standard deviations** of the mean

Why Standard Normal Distribution?

Suppose we literally want to compare apples to oranges. Specifically, let's compare their weights. Imagine that we have an apple that weighs 110 grams and an orange that weighs 100 grams.
If we compare the raw values, it's easy to see that the apple weighs more than the orange. However, let's compare their standard scores. To do this, we'll need to know the properties of the weight distributions for apples and oranges. Assume that the weights of apples and oranges follow a normal distribution with the following parameter values:



Apples	Oranges	
Mean weight grams	100	140
Standard Deviation	15	25

Now we'll calculate the Z-scores:

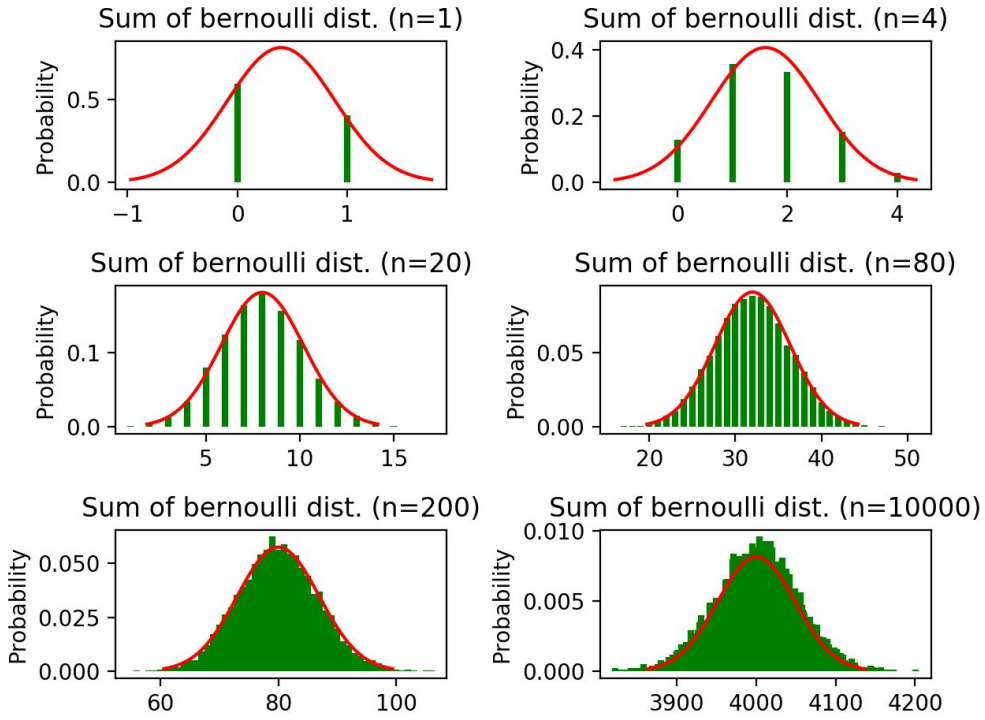
- Apple = (110-100) / 15 = 0.667
- Orange = (100-140) / 25 = -1.6

The Z-score for the apple (0.667) is positive, which means that our apple weighs more than the average apple. It's not an extreme value by any means, but it is above average for apples. On the other hand, the orange has fairly negative Z-score (-1.6). It's pretty far below the mean weight for oranges. I've placed these Z-values in the standard normal distribution below.

Central limit theorem

- Concept – if we draw infinite samples of size n from a distribution and plot the means of those samples, we get sampling distribution of the mean. This is theoretical concept as drawing infinite samples is not practical.

- The sampling distribution of the means of those samples will become approximately normally distributed with mean μ and standard deviation σ / (square root of) N as the sample size (N) and the number of samples taken become larger, irrespective of the shape of the population distribution.
- The mean value of the distribution will be close estimate pf the population mean μ.



HYPOTHESIS TESTING

- **Null Hypothesis (H) –** It is a statement that is commonly accepted or is considered to be the status quo. It is assumed that the observed result is due to the chance of factor. It is denoted by H. If it is a test of means then we say that **H: $\mu_1 = \mu_2$** , which states that there is no significant difference in the 2 population means.

- **Alternate Hypothesis(H1 or Ha) –** As previously mentioned that Null Hypothesis and Alternate Hypothesis are mutually exclusive statements. So if the Null Hypothesis is commonly accepted facts then the Alternate Hypothesis is a real fact-based on observation from the sample data. It is denoted by H1 or Ha. If it is a test of means then we say that H1 : $\mu 1 \neq \mu 2$, which states that there is a significant difference in 2 population means.

Fundamentals of Hypothesis Testing

•Let’s take an example to understand the concept of Hypothesis Testing. A person is on trial for a criminal offense and the judge needs to provide a verdict on his case. Now, there are four possible combinations in such a case:

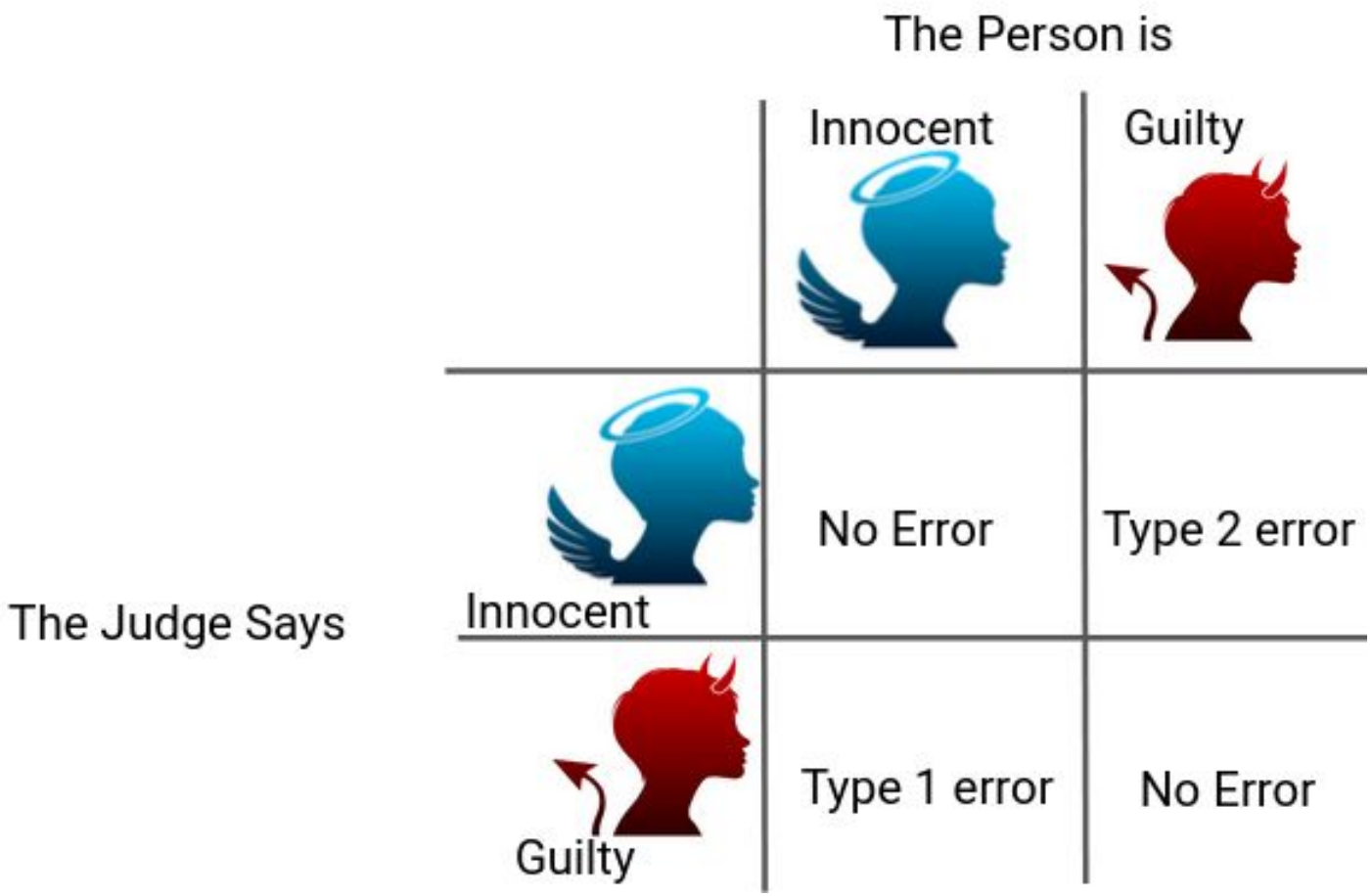
•First Case: The person is innocent and the judge identifies the person as innocent

•Second Case: The person is innocent and the judge identifies the person as guilty

•Third Case: The person is guilty and the judge identifies the person as innocent

•Fourth Case: The person is guilty and the judge identifies the person as guilty

•



•As you can clearly see, there can be two types of error in the judgment – Type 1 error, when the verdict is against the person while he was innocent and Type 2 error, when the verdict is in favor of Person while he was guilty

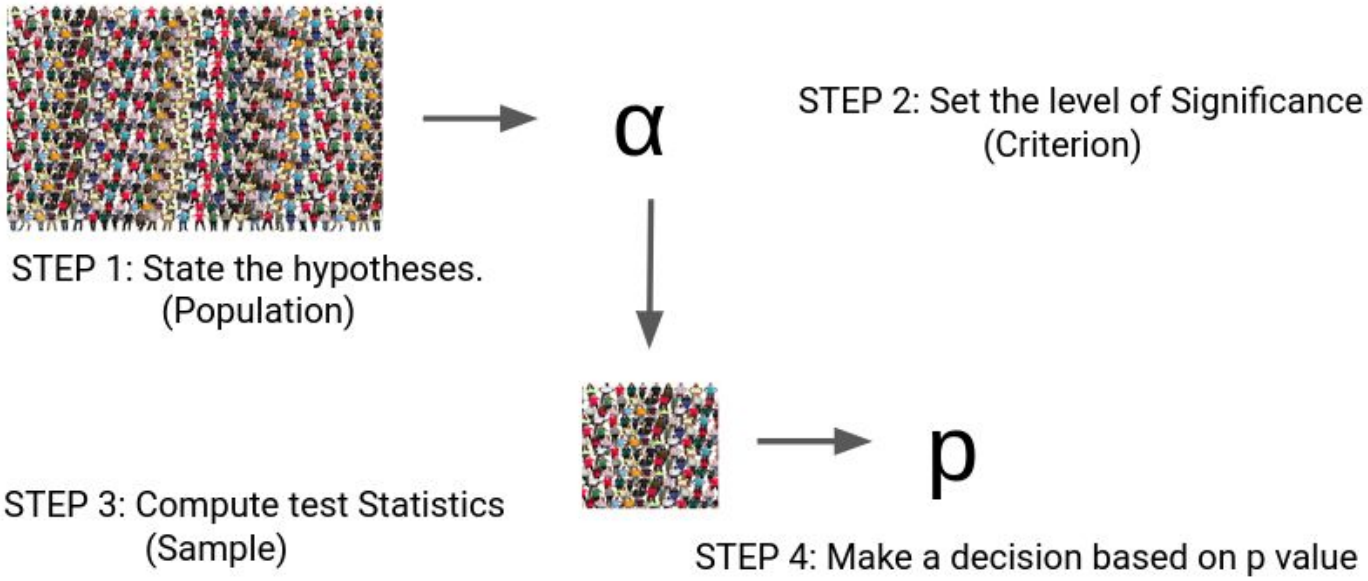
•According to the Presumption of Innocence, the person is considered innocent until proven guilty. That means the judge must find the evidence which convinces him "beyond a reasonable doubt". This phenomenon of **"Beyond a reasonable doubt"** can be understood as **Probability (Judge Decided Guilty | Person is Innocent) should be small.**

•The basic concepts of Hypothesis Testing are actually quite analogous to this situation.

•We consider **the Null Hypothesis** to be true until we find strong evidence against it. Then. we accept the **Alternate Hypothesis**. We also determine the **Significance Level (α)** which can be understood as the Probability of (Judge Decided Guilty | Person is Innocent) in the previous example. Thus, if α is smaller, it will require more evidence to reject the Null Hypothesis. Don't worry, we'll cover all of this using a case study later.

•

Steps to Perform Hypothesis testing

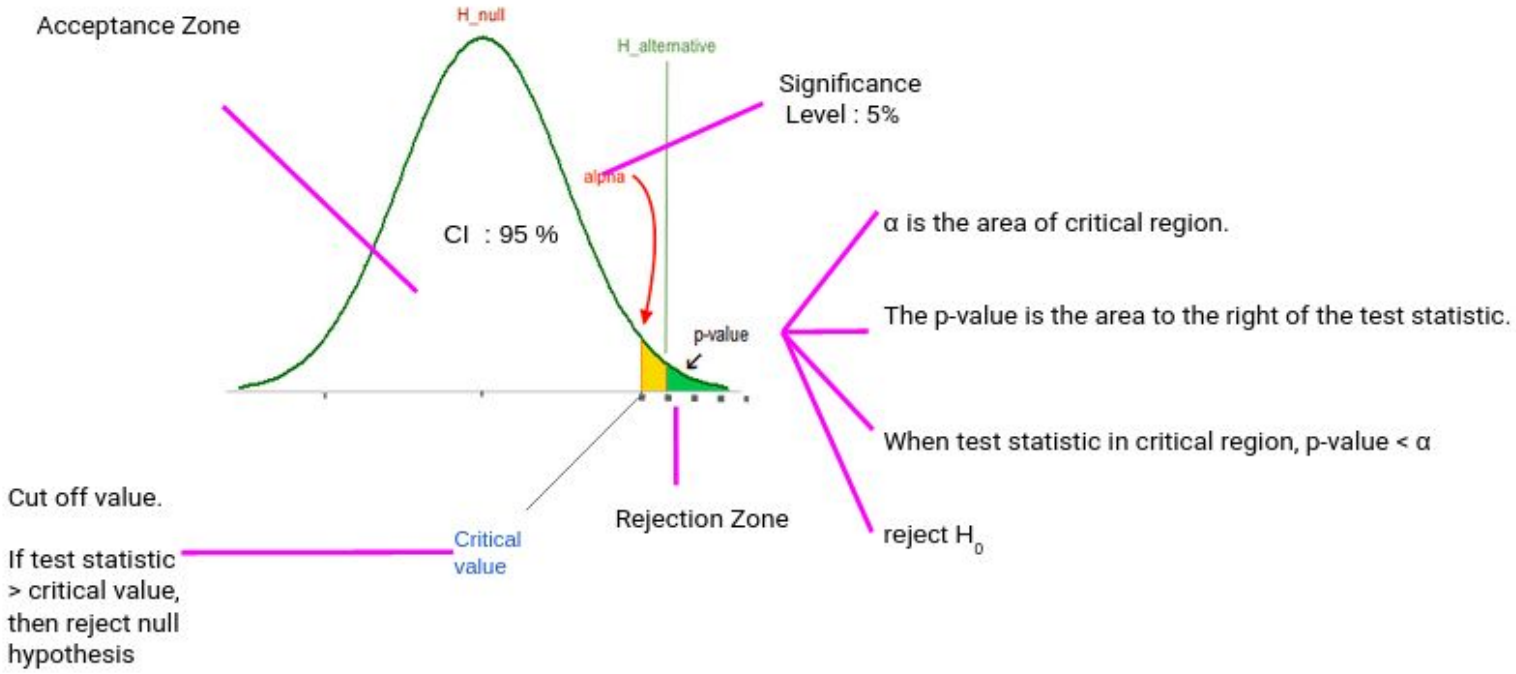


Steps 1 to 3 are quite self-explanatory but on what basis can we make a decision in step 4? What does this p-value indicate?
We can understand this p-value as the measurement of the Defense Attorney's argument. If the p-value is less than α , we reject the Null Hypothesis or if the p-value is greater than α, we fail to reject the Null Hypothesis.

- Set the Hypothesis
- Set the Significance Level, Criteria for a decision
- Compute the test statistics
- Make a decision

Critical Value, p-value

Critical Value *is the cut off value between Acceptance Zone and Rejection Zone. We compare our test score to the critical value and if the test score is greater than the critical value, that means our test score lies in the Rejection Zone and we reject the Null Hypothesis. On the opposite side, if the test score is less than the Critical Value, that means the test score lies in the Acceptance Zone and we fail to reject the null Hypothesis.*

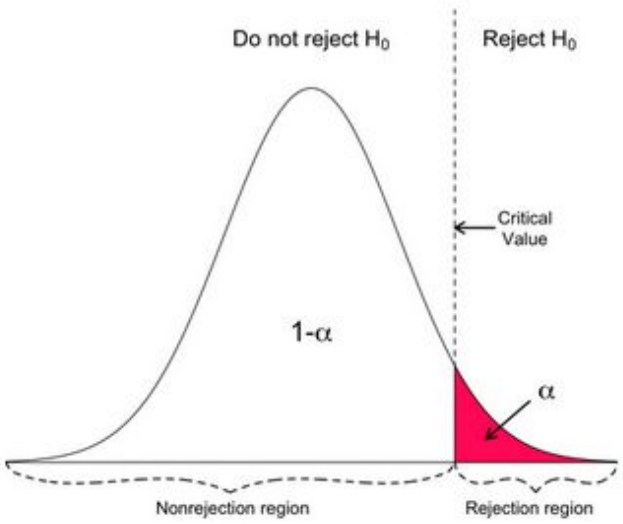


Typically, we set the Significance level at 10%, 5%, or 1%. If our test score lies in the Acceptance Zone we fail to reject the Null Hypothesis. If our test score lies in the critical zone, we reject the Null Hypothesis and accept the Alternate Hypothesis.

Hypothesis Test		Test Statistic
Z test	One sample Z test	Z - Statistics
	Two sample Z test	
	Two Proportions Z test	
T-tests	One sample T test	T - Statistics
	Two Sample T test	
	Paired T test	
Chi Square	Chi Square Fit test	Chi Square Statistics
	Chi Square test of independence	
F test	One Way ANOVA	F-Statistics
	Poc Host test	
	Factorial Analysis of Variance	
	Repeated measure of variance	
	Analysis of covariance	

Directional Hypothesis

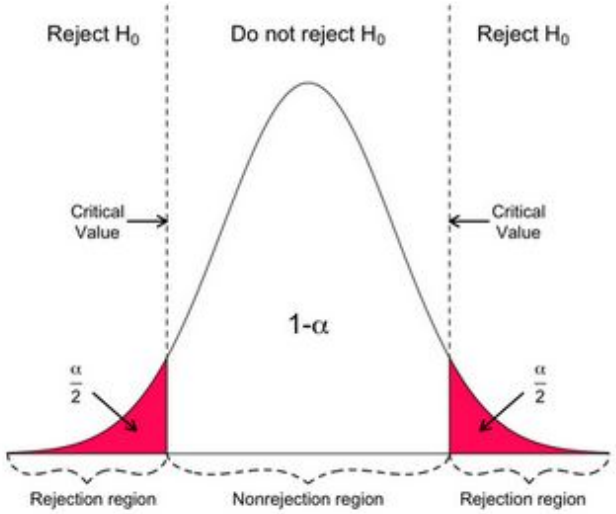
In the Directional Hypothesis, the null hypothesis is rejected if the test score is too large (for right-tailed and too small for left tailed). Thus, the rejection region for such a test consists of one part, which is right from the center.



Ex. Females will score significantly higher than Males on IQ score

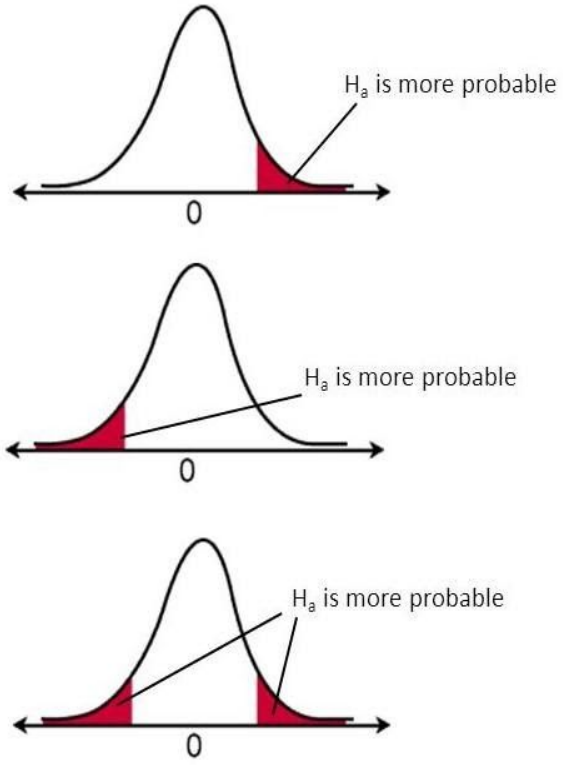
Non-Directional Hypothesis

In a Non-Directional Hypothesis test, the Null Hypothesis is rejected if the test score is either too small or too large. Thus, the rejection region for such a test consists of two parts: one on the left and one on the right.



Ex. Females will score significantly different than Males on IQ score

- **Critical Region –** The critical region is defined as the region of values in distribution that leads to the rejection of the null hypothesis at some given probability level.
- **One-Tailed Test –** A one-tailed test is a statistical hypothesis test in which the critical area of distribution is either greater than or less than a certain value, but can't be both. For this the alternate hypothesis formulation is $\mathbf{H_1 : \mu_1 > \mu_2}$ **or** $\mathbf{H_1 : \mu_1 < \mu_2}$.
- **Two-Tailed Test –** A two-tailed test is a statistical hypothesis test in which the critical area of distribution is on either of the sides. It tests whether the sample means of 2 or more populations are unequal (in the test of means). For this alternate hypothesis, the formulation is $\mathbf{H_1 : \mu_1 \neq \mu_2}$.



Right-tail test

$H_a: \mu > \text{value}$

Left-tail test

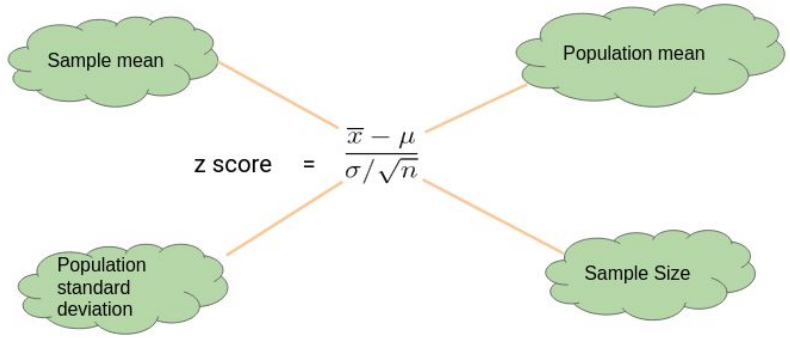
$H_a: \mu < \text{value}$

Two-tail test

$H_a: \mu \neq \text{value}$

What is the Z Test?

z tests are a statistical way of testing a hypothesis when either:
•We know the population variance, or
•We do not know the population variance but our sample size is large n ≥ 30
If we have a sample size of less than 30 and do not know the population variance, then we must use a t-test.



Here's an Example to Understand a One Sample Z Test

Score
650
730
510
670
480
800
690
530
590
620
710
670
640
780
650
490
800
600
510
700

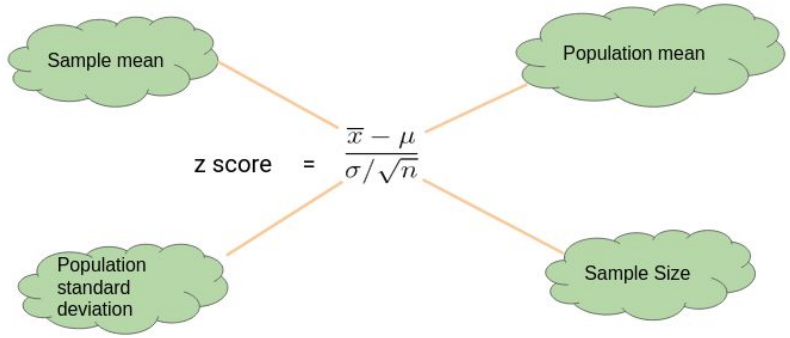
Let's say we need to determine if girls on average score higher than 600 in the exam. We have the information that the standard deviation for girls' scores is 100. So, we collect the data of 20 girls by using random samples and record their marks. Finally, we also set our α value (significance level) to be 0.05.

Null Hypothesis: Girls on avg score less than 600, m1 < 600

Alternate Hypothesis: Girls on avg score higher than 600 , m1 > 600

What is the Z Test?

z tests are a statistical way of testing a hypothesis when either:
•We know the population variance, or
•We do not know the population variance but our sample size is large n ≥ 30
If we have a sample size of less than 30 and do not know the population variance, then we must use a t-test.



Here's an Example to Understand a One Sample Z Test

Let's say we need to determine if girls on average score higher than 600 in the exam. We have the information that the standard deviation for girls' scores is 100. So, we collect the data of 20 girls by using random samples and record their marks. Finally, we also set our α value (significance level) to be 0.05.

Score
650
730
510
670
480
800
690
530
590
620
710
670
640
780
650
490
800
600
510
700

Solution:

In this example:
•Mean Score for Girls is 641
•The size of the sample is 20
•The population mean is 600
•Standard Deviation for Population is 100

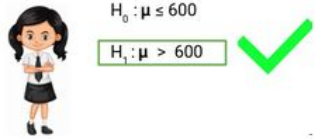
$$\text{z score} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{641 - 600}{100/\sqrt{20}} = 1.8336$$

p value = .033357.

Critical Value = 1.645

Z score > Critical Value

P value < 0.05



Since the P-value is less than 0.05, we can reject the null hypothesis and conclude based on our result that Girls on average scored higher than 600.

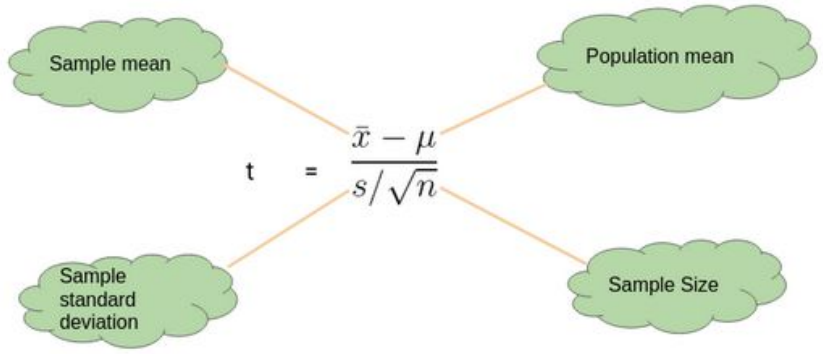
What is the t-Test?

t-tests are a statistical way of testing a hypothesis when:

- We do not know the population variance
- Our sample size is small, n < 30

One-Sample t-Test

We perform a One-Sample t-test when we want to **compare a sample mean with the population mean**. The difference from the Z Test is that we do **not have the information on Population Variance** here. We use the **sample standard deviation** instead of population standard deviation in this case.



Examples

Here's an Example to Understand a One Sample t-Test

Let's say we want to determine if on average girls score more than 600 in the exam. We do not have the information related to variance (or standard deviation) for girls' scores. To a perform t-test, we randomly collect the data of 10 girls with their marks and choose our α value (significance level) to be 0.05 for Hypothesis Testing.

Girls_Score
587
602
627
610
619
622
605
608
596
592

In this example:

- Mean Score for Girls is 606.8
- The size of the sample is 10
- The population mean is 600
- Standard Deviation for the sample is 13.14

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

$$= \frac{606.8 - 600}{13.14/\sqrt{10}}$$

$$= 1.64$$

Critical Value = 1.833

t score < Critical Value

P value = 0.0678

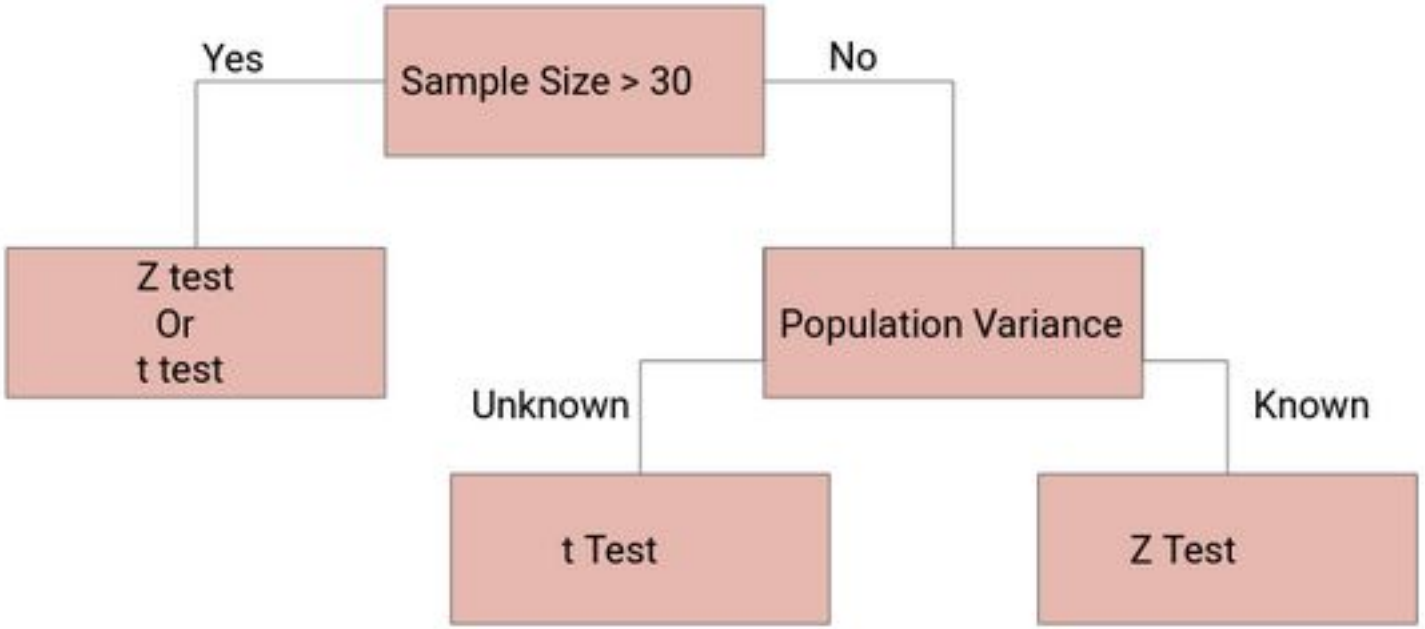
P value > 0.05

$H_0 : \mu \leq 600$

$H_1 : \mu > 600$

Our **P-value is greater than 0.05 thus we fail to reject the null hypothesis** and don't have enough evidence to support the hypothesis that on average, girls score more than 600 in the exam.

Deciding between Z Test and T-Test



Case Study: Hypothesis Testing for Coronavirus using Python

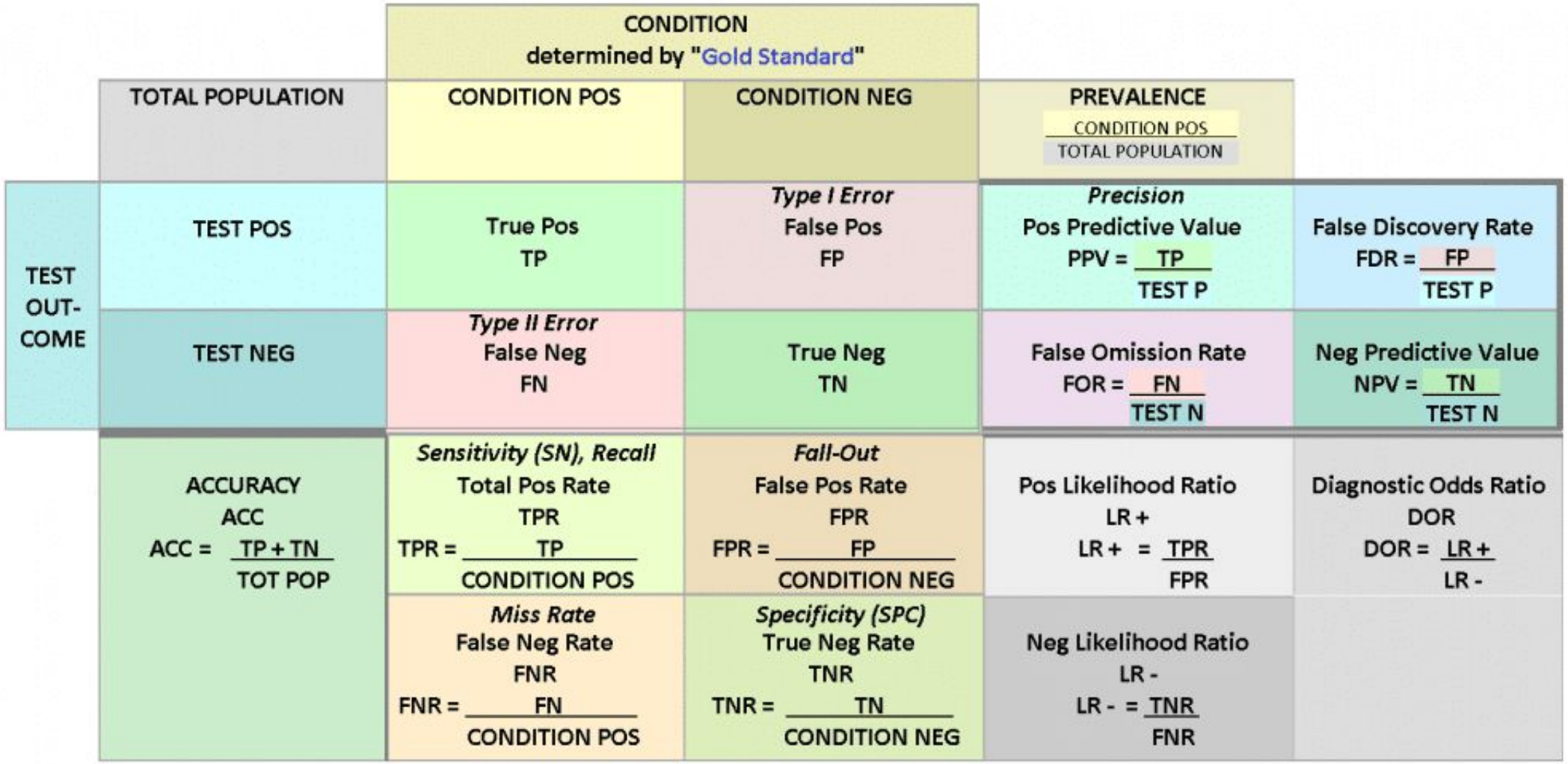
Now let's implement the Two-Sample Z test for a coronavirus dataset. Let's put our theoretical knowledge into practice and see how well we can do. You can download the dataset here.
This dataset has been taken from **John Hopkin's repository** and you can find the link here for it.

Type-I error – This error occurs when insights drawn from sample data lead to rejection of the null hypothesis even when it is true. This error could be controlled as it has direct bearing with a **level of significance.**

Type-II error – This error occurs when insights from sample data result in failing to reject the null hypothesis although it is false.

Level of Significance – It is the probability of making type I error and is denoted by **α.** It is the maximum probability of making type I error. As per standard for 95% confidence level value of alpha is 0.05. This means that there is a 5% probability of making type I error or rejecting the null hypothesis even when it is true.

p-value – It is a statistical concept that is used from hypothesis testing to regression to tree models and much more. It is an integral part of data science. If the p-value is high there are higher chances of the null hypothesis being true and if the p-value is low then it is more likely to reject the null hypothesis. The Standard p-value is equal to alpha and is used for checking statistical p-value against it and making the decision.


CONDITION
determined by "Gold Standard"
TOTAL POPULATION
CONDITION POS
CONDITION NEG
PREVALENCE
CONDITION POS / TOTAL POPULATION
TEST OUT-COME
TEST POS
True Pos
TP
Type I Error
False Pos
FP
Precision
Pos Predictive Value
PPV = TP / TEST P
False Discovery Rate
FDR = FP / TEST P
TEST NEG
Type II Error
False Neg
FN
True Neg
TN
False Omission Rate
FOR = FN / TEST N
Neg Predictive Value
NPV = TN / TEST N
ACCURACY
ACC
ACC = TP + TN / TOT POP
Sensitivity (SN), Recall
Total Pos Rate
TPR
TPR = TP / CONDITION POS
Fall-Out
False Pos Rate
FPR
FPR = FP / CONDITION NEG
Pos Likelihood Ratio
LR +
LR + = TPR / FPR
Diagnostic Odds Ratio
DOR
DOR = LR + / LR -
Miss Rate
False Neg Rate
FNR
FNR = FN / CONDITION POS
Specificity (SPC)
True Neg Rate
TNR
TNR = TN / CONDITION NEG
Neg Likelihood Ratio
LR -
LR - = TNR / FNR