Statistics for Data Science

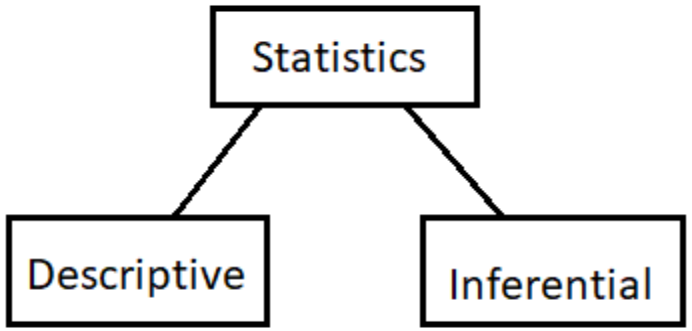
Definition

Statistics is the science, or a branch of mathematics, that involves collecting, classifying, analyzing, interpreting, and presenting numerical facts and data. It is especially handy when dealing with populations too numerous and extensive for specific, detailed measurements. Statistics are crucial for drawing general conclusions relating to a dataset from a data sample.

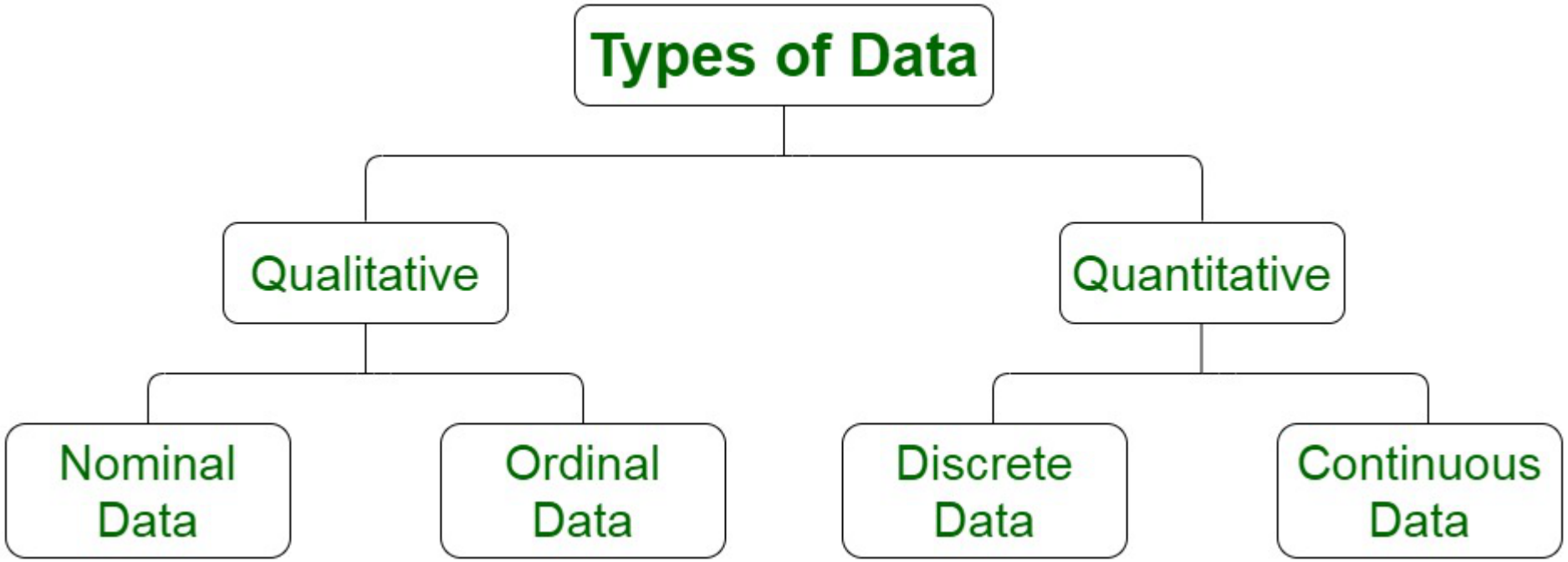
Types of Statistics

There are two types of Statistics:

1. Descriptive Statistics
2. Inferential Statistics



Types of Data



Let’s explore all types of data

Quantitative Data

1. Discrete Data
    - It can take only discrete values. Discrete information contains only a finite number of possible values. Those values cannot be subdivided meaningfully. Here, things can be counted in whole numbers.
    - **Example:** Number of students in the class, Number of bank accounts
2. Continuous data
    - It represents measurements and therefore their values can’t be counted but they can be measured.
    - **Example**: Height of a person (which you can describe by using intervals on the real number line), Average Rainfall, Body Temperature

Qualitative Data/Categorical Data

1. Nominal Data
    - Nominal values represent discrete units and are used to label variables that have no quantitative value. Just think of them as “labels.” Note that nominal data that has no order. Therefore, if you would change the order of its values, the meaning would not change
    - **Example**: Gender Type (Male, Female or Others), Language spoken by an individual (English, Spanish, French, Hindi, or Others)
2. Ordinal Data
    - Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that its ordering matters
    - Example: Student’s performance in the exam(Outstanding, Good, Average, Unsatisfactory, Failed). You can associate a rank or an order with each and every labels, i.e. Outstanding (1) , Good (2) and so on.

Example of all types of data in a tabular form

Let's take an example of 'Student' table below

Age	Height	Sex	Academic Performance
21	5.7	Male	Average
20	5.5	Female	Good
23	5.9	Male	Outstanding

From the above example, we can see all four types of data. 'Age' is Discrete, 'Height' is Continuous, 'Sex' is Nominal and 'Academic Performance' is Ordinal data

Sample Data & Population Data

A **population** is the entire group that you want to draw conclusions about.

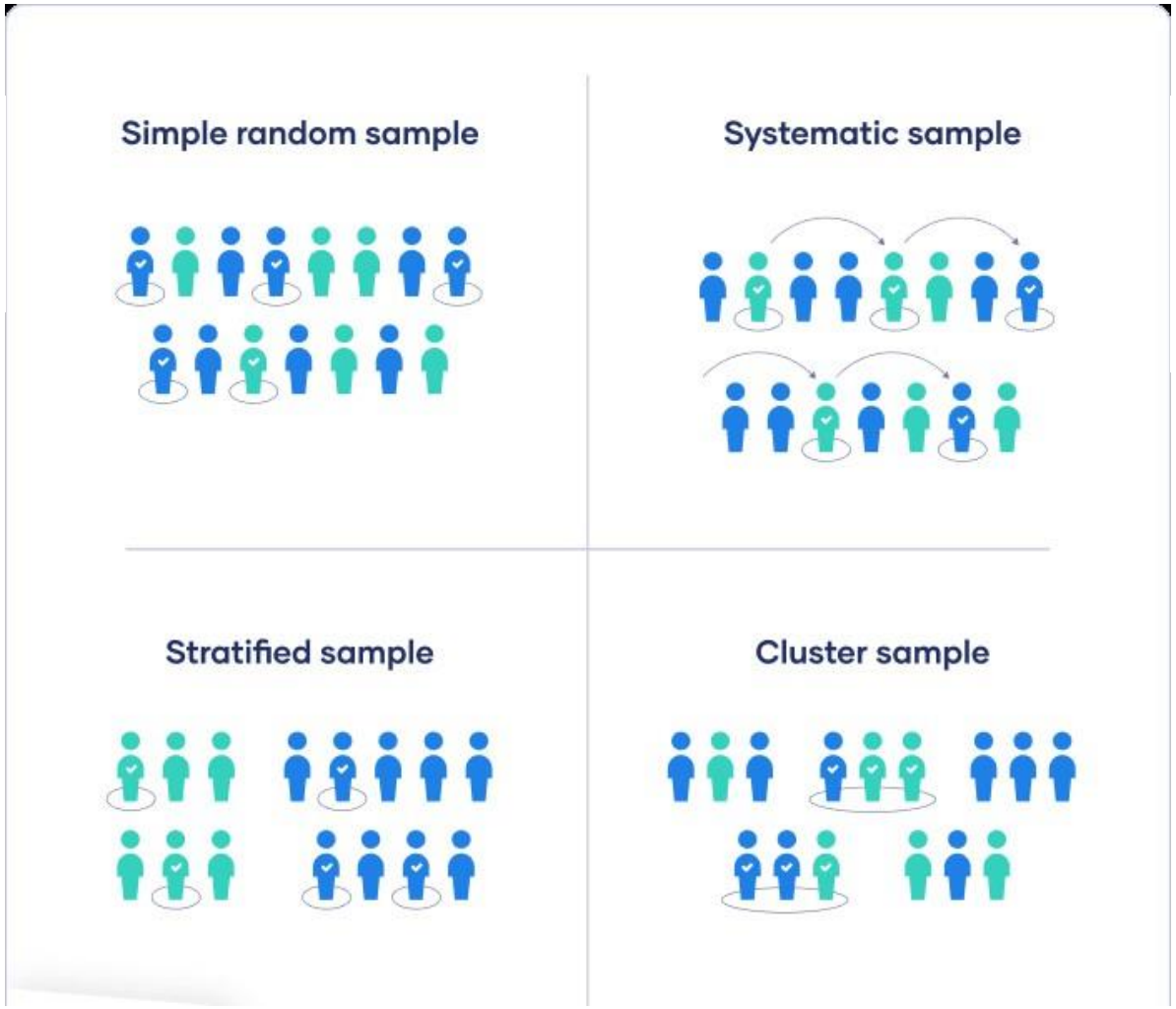
A **sample** is the specific group that you will collect data from. The size of the sample is always less than the total size of the population.

Sampling Techniques:

- Simple Random Sampling
- Systematic Sampling
- Stratified Sampling
- Cluster Sampling

Most Popular Sampling Techniques

Every member of the population has an equal chance of being selected



Similar to simple random Sampling, but individuals are chosen at regular intervals

Divide the population into subgroups (called strata) based on the relevant characteristic (e.g., gender identity, age). Then using random or systematic sampling to select a sample from each subgroup

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have same attributes to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups

Descriptive Statistics

- Descriptive statistics describe, show, and summarize the basic features of a dataset found in a given study, presented in a summary that describes the data sample and its measurements. It helps analysts to understand the data better.
- Descriptive statistics represent the available data sample and do not include theories, inferences, probabilities, or conclusions. That's a job for inferential statistics.

Topics under descriptive statistics:

1. Measures of central tendency
2. Measures of variability
3. Distribution (Also Called Frequency Distribution)

Let's start with one topic at a time

Measures of Central Tendency

There are three fundamental concepts under this topic:

1. Mean
2. Median
3. Mode

Let’s have a look one concept at a time

Mean

The arithmetic mean of a raw data is obtained by adding all the values of the variables and dividing the sum by total number of values that are added.

Sample mean is denoted by $\bar{x}$, whereas population mean is denoted by μ

$\bar{x} = \sum x_i / n = (x_1+x_2+x_3+\ldots\ldots+x_n)/n$ (n is the sample size)

$\mu = \sum X / N = (X_1+X_2+X_3+\ldots\ldots+X_n)/N$ (N is the population size)

Mean manages to provide a central value of the data distribution, however, there is a big disadvantage of this technique. If we have **outliers** in our dataset, the mean of the distribution is heavily impacted by those extreme values. Let's take an example to understand this-

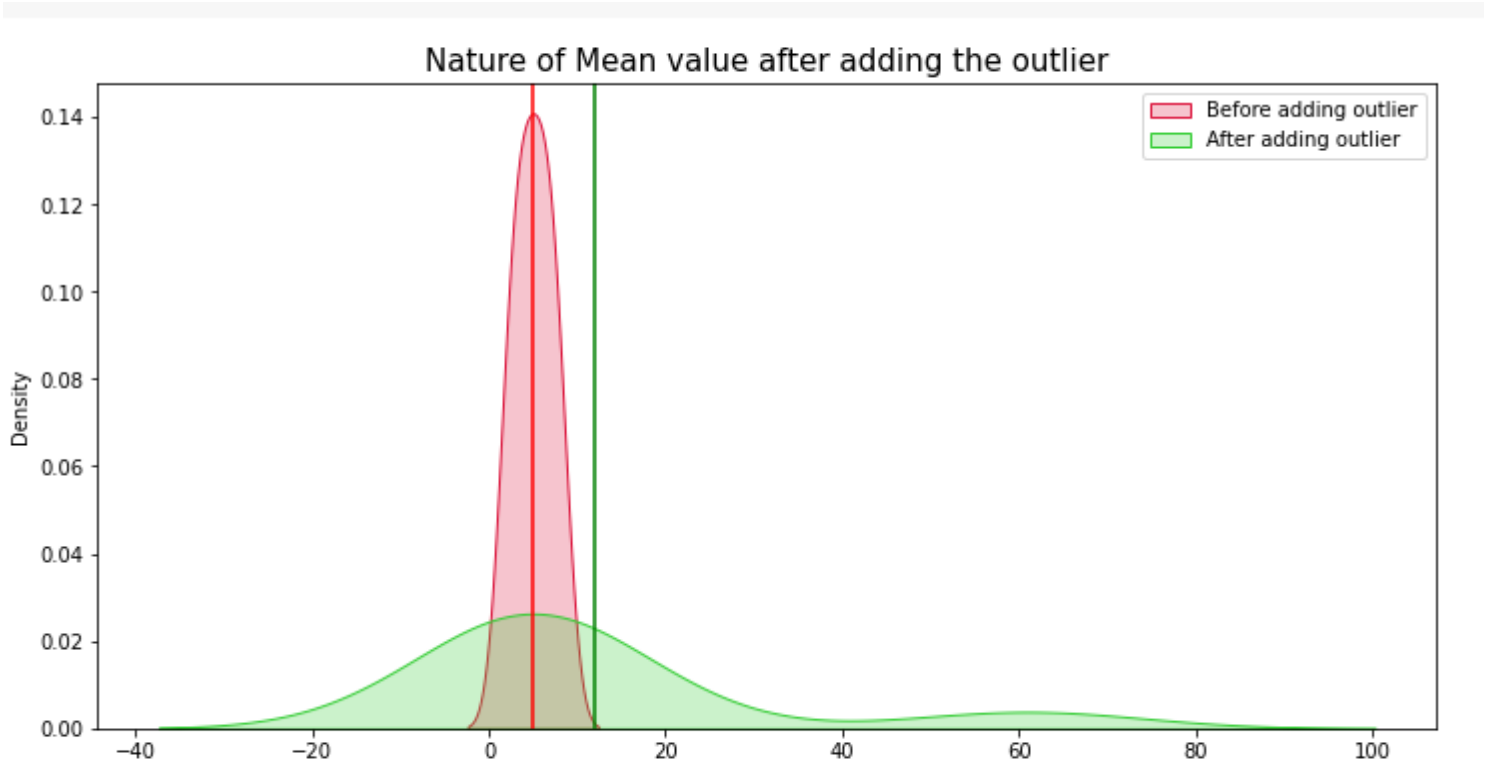
data=[4,5,3,6,8,2,7] , if we calculate the mean of this list, the result will be 5

Now, let's add an outlier.

data=[4,5,3,6,8,2,7,**61**] , here 61 is an outlier, the resulting mean will be 12

Mean Contd.

So, we can see that one outlier has changed the mean from 5 to 12.

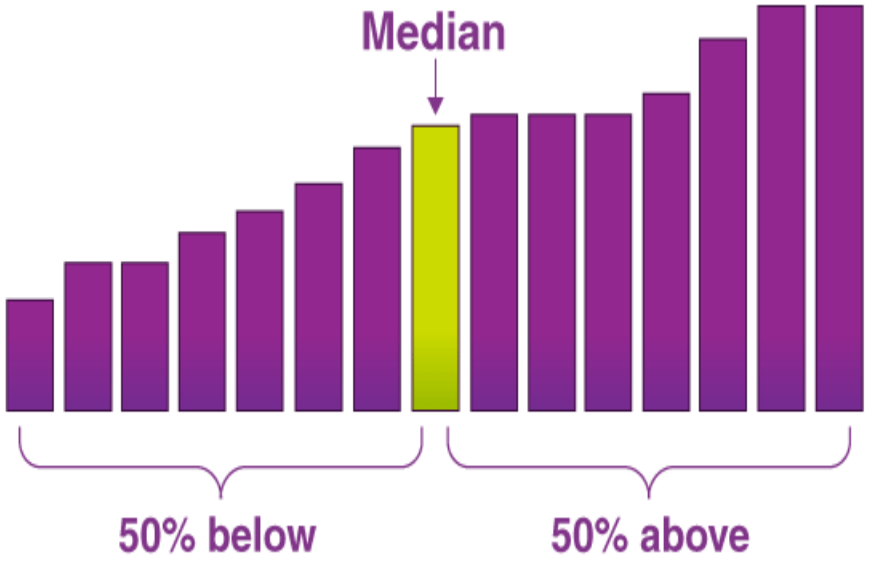


Red vertical line represents the mean before adding the outlier, green vertical line represents the mean after adding the outlier.

Median

The median of a set of data is the middlemost number or centre value in the set. The median is also the number that is halfway into the set.

To find the median, the data should be arranged first in order of least to greatest or greatest to the least value



- For odd number of observations

  Median= $((n+1)/2)^{th}$ term

- For even number of observations

  Median= Mean of $(n/2)^{th}$ and $((n/2)+1)^{th}$ terms

(n is number of observations)

Ex: [2,4,5,7,8] => Median is $((n+1)/2)^{th}$ term => $((5+1)/2)^{th}$ term => 5

Ex: [2,3,6,8] => Median is mean of $(n/2)^{th}$ and $((n/2)+1)^{th}$ terms =>(3+6)/2=>4.5

Median Contd.

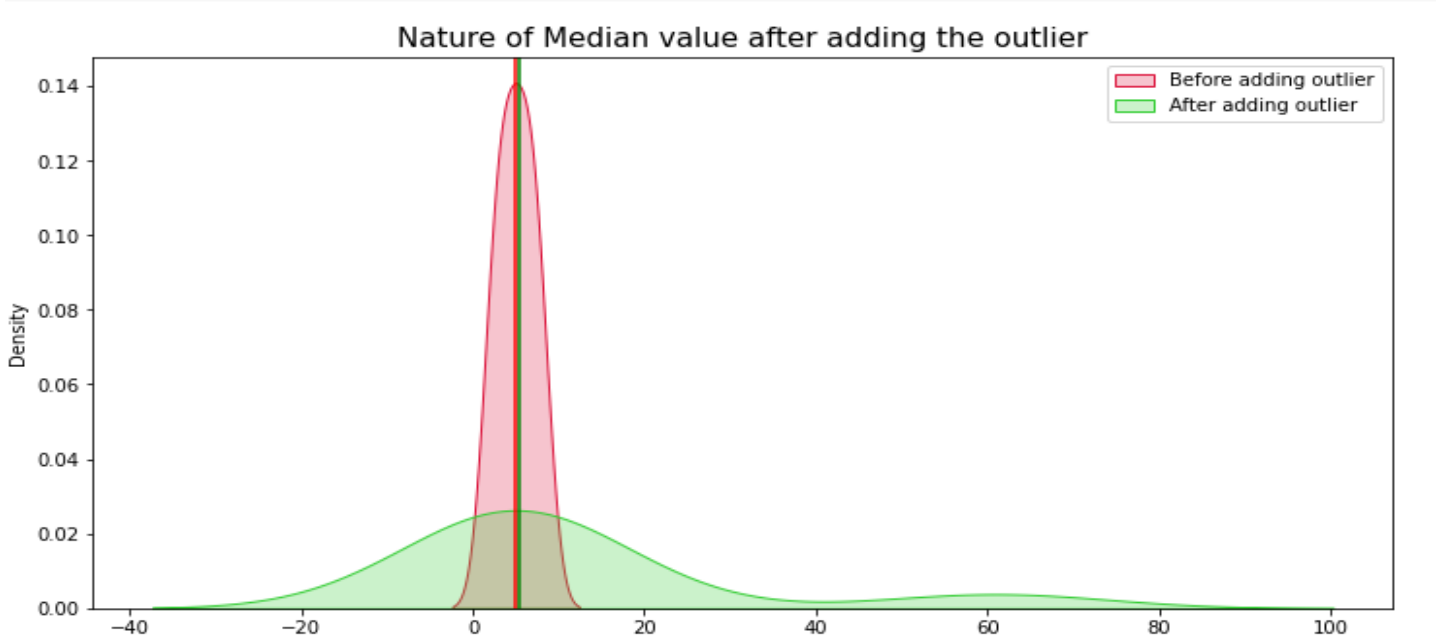
Unlike mean, median does not get affected by the outliers. Let's take an example to understand this.

data=[2,3,4,6,8], median is 4

Let's add an outlier, data=[2,3,4,6,8,**78**], and the median is (4+6)/2=5

So, the change in median is very less as compared to mean.

From the diagram, we can see that the red vertical line is almost coinciding with the green vertical line. Both Vertical lines are representing median



Mode

- A mode is defined as the value that has a higher frequency in a given set of values. It is the value that appears the most number of times.
- **Example 1**: In the given set of data: 2, 4, 5, 5, 6, 7, the mode of the data set is 5 since it has appeared in the set twice.
- Example 2: In the given set of data: 2, 4, 5, 5, 6, 7, 8, 8, the modes of the data set are 5 and 8 since both the numbers have appeared in the set twice

Python Implementation of Mean, Median, and, Mode:

Please follow this GitHub link (this is a hyperlink) to find the implementation

Measures of Variability

Variability is most commonly measured with the following descriptive statistics:

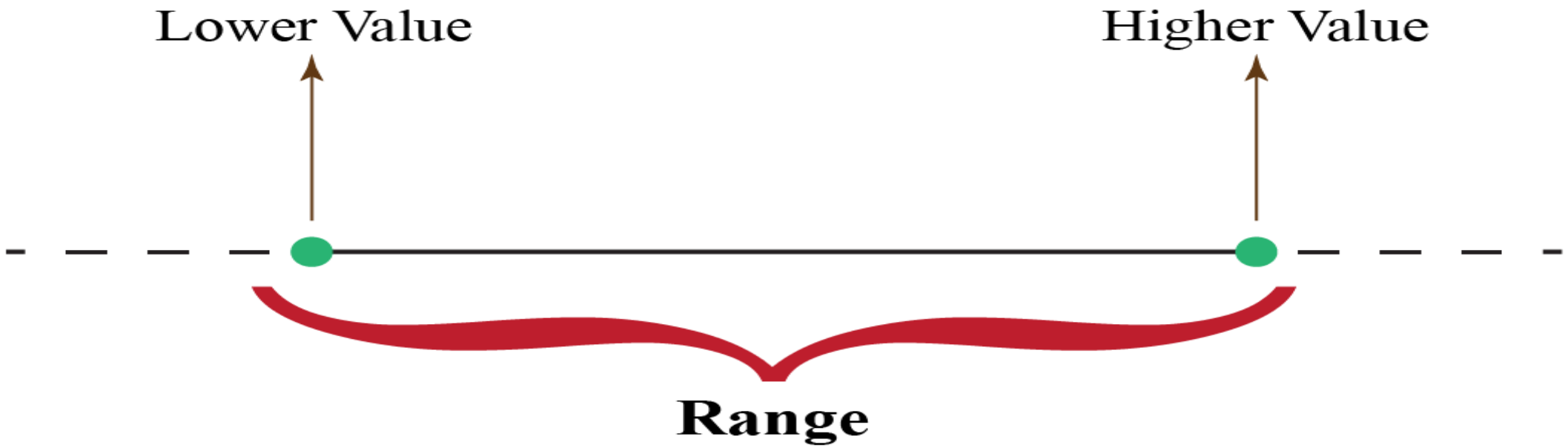
1. Range
2. Interquartile range
3. Standard deviation
4. Variance

Let’s have a look once concept at a time

Range

In statistics, the **range** is the spread of your data from the lowest to the highest value in the distribution

Ex: {5,2,7,9,1,15,8,-1,10}. For this distribution, the lowest value is -1 and the highest value is 15. So the Range of this distribution will be (15-(-1))=16



Variance

- The **variance** is a measure of variability. It is calculated by taking the average of squared deviations from the mean.
- Variance tells you the degree of spread in your data set. The more spread the data, the larger the variance is in relation to the mean.

- **Population variance**
- When you have collected data from every member of the population that you're interested in, you can get an exact value for population variance.
- The **population variance** formula looks like this: $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{\mu})^2}{N}$

N: The number of values in population, $\overline{\mu}$: Population mean

Variance Contd.

Sample variance: $s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{x})^2}{n-1}$

n: The number of values in sample, $\overline{x}$: Sample mean

- With samples, we use *n* – 1 in the formula because using n would give us a biased estimate that consistently underestimates variability. The sample variance would tend to be lower than the real variance of the population.
- Reducing the sample *n* to *n* – 1 makes the variance artificially large, giving you an unbiased estimate of variability: it is better to overestimate rather than underestimate variability in samples.

Variance Example

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

x	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
1	2.83	1 – 2.83 = (-1.83)	(-1.83)^2 = 3.35
2	2.83	2 – 2.83 = (-0.83)	(-0.83)^2 = 0.69
2	2.83	2 – 2.83 = (-0.83)	(-0.83)^2 = 0.69
3	2.83	3 – 2.83 = (0.17)	(0.17)^2 = 0.03
4	2.83	4 – 2.83 = (1.17)	(1.17)^2 = 1.37
5	2.83	5 – 2.83 = (2.17)	(2.17)^2 = 4.71

3.35 + 0.69 + 0.69 + 0.03 + 1.37 + 4.71 = 10.84

$$s^2 = \frac{10.84}{6 - 1} = 2.17$$

Standard Deviation

A standard deviation (or σ) is a measure of how dispersed the data is in relation to the mean. Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out.

$s = \sqrt{\left(\frac{\sum_{i=1}^{n}(X_i - \overline{x})^2}{n-1}\right)}$ (Sample SD)

$\sigma = \sqrt{\left(\frac{\sum_{i=1}^{N}(X_i - \overline{\mu})^2}{N}\right)}$ (Population SD)

- **Importance of SD:** Standard deviation is a useful measure of spread for normal distributions. In normal distributions, data is symmetrically distributed with no skew. Most values cluster around a central region, with values tapering off as they go further away from the center. The standard deviation tells you how spread out from the center of the distribution your data is on average.
- Many scientific variables follow normal distributions, including height, standardized test scores, or job satisfaction ratings

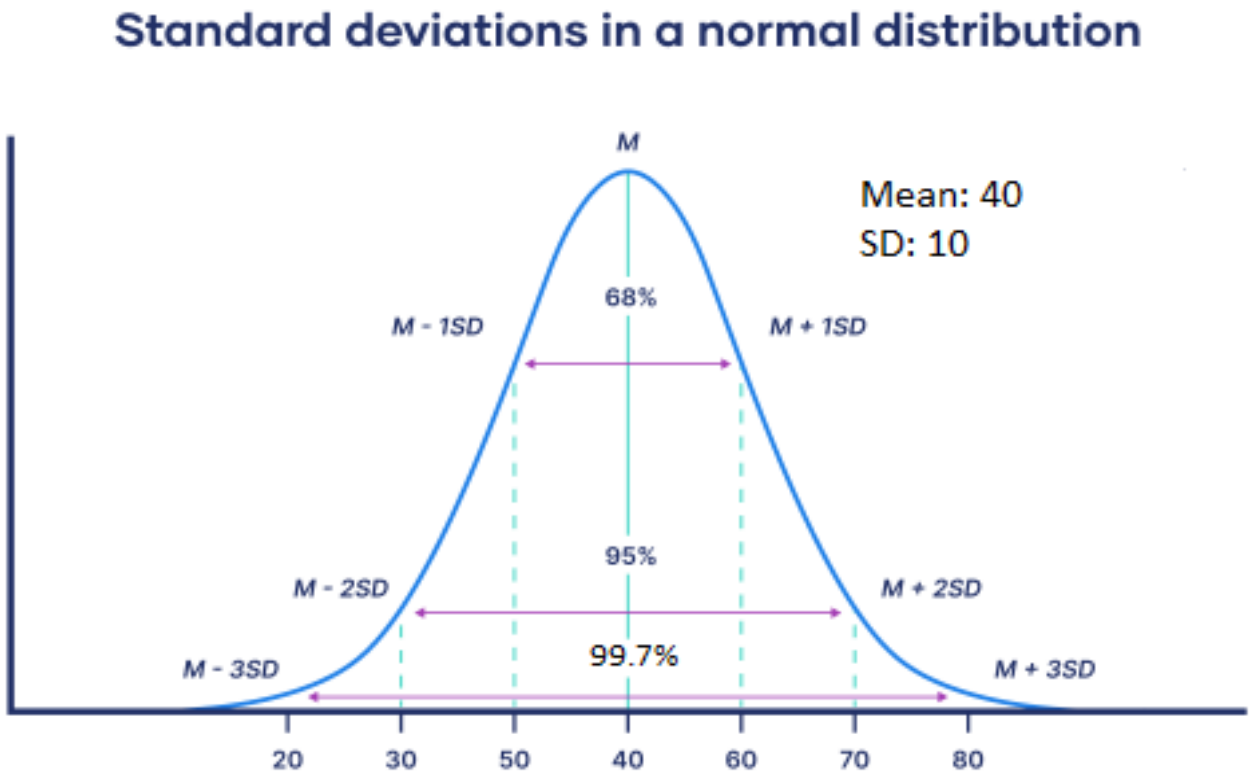
SD Contd.

The standard deviation reflects the dispersion of the distribution. The curve with the lowest standard deviation has a high peak and a small spread, while the curve with the highest standard deviation is more flat and widespread.

Empirical rule: The standard deviation and the mean together can tell you where most of the values in your frequency distribution lie if they follow a normal distribution.

- Around 68% of scores are within 1 standard deviation of the mean
- Around 95% of scores are within 2 standard deviations of the mean
- Around 99.7% of scores are within 3 standard deviations of the mean.

SD Contd.


Standard deviations in a normal distribution
M
Mean: 40
SD: 10
68%
M - 1SD
M + 1SD
95%
M - 2SD
M + 2SD
99.7%
M - 3SD
M + 3SD
20
30
50
40
60
70
80

Interquartile Range

5 number summary: The five number summary provides this information using various descriptive statistics. These statistics are all order statistics—each one describes where a particular value falls in the distribution. The five statistics in this summary are the following, from highest to lowest data values:

- Highest value in the dataset.
- Third quartile (Q3)—greater than 75% of the values in the dataset
- Median or second quartile (Q2)—splits the dataset in half.
- First quartile (Q1)—greater than 25% of the values.
- Lowest value in the dataset.

IQR = Q3 – Q1

In order to remove the outliers, we can define the upper bound and the lower bound of the distribution as following:

IQR Contd.

Upper Bound = Q1 – (1.5*IQR)

Lower Bound = Q3 + (1.5*IQR)

The data points which are greater than Upper Bound or less than Lower Bound, will be considered as outliers and we will discard those values in our analysis.

Let's take an example to understand this concept

data_points = [4,6,2,8,9,7,-1,36,-28]

75^{th} percentile (Q3) = 8

25^{th} percentile (Q1) = 2

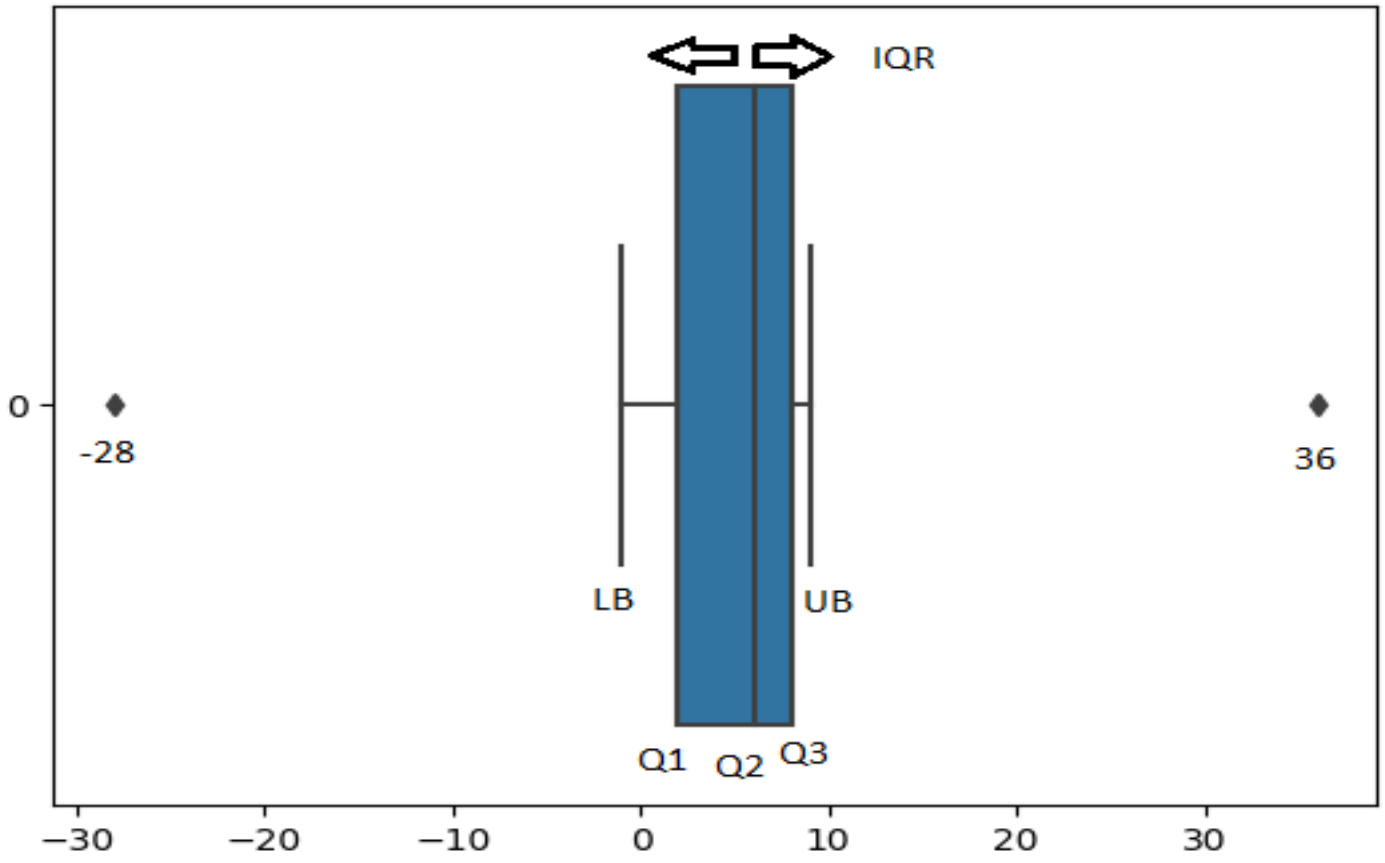
IQR = 8-2 = 6

Upper Bound = 8+(1.5*6) = 17

Lower Bound = 2-(1.5*6) = -7

IQR Contd.

Outliers will be [36,-28]



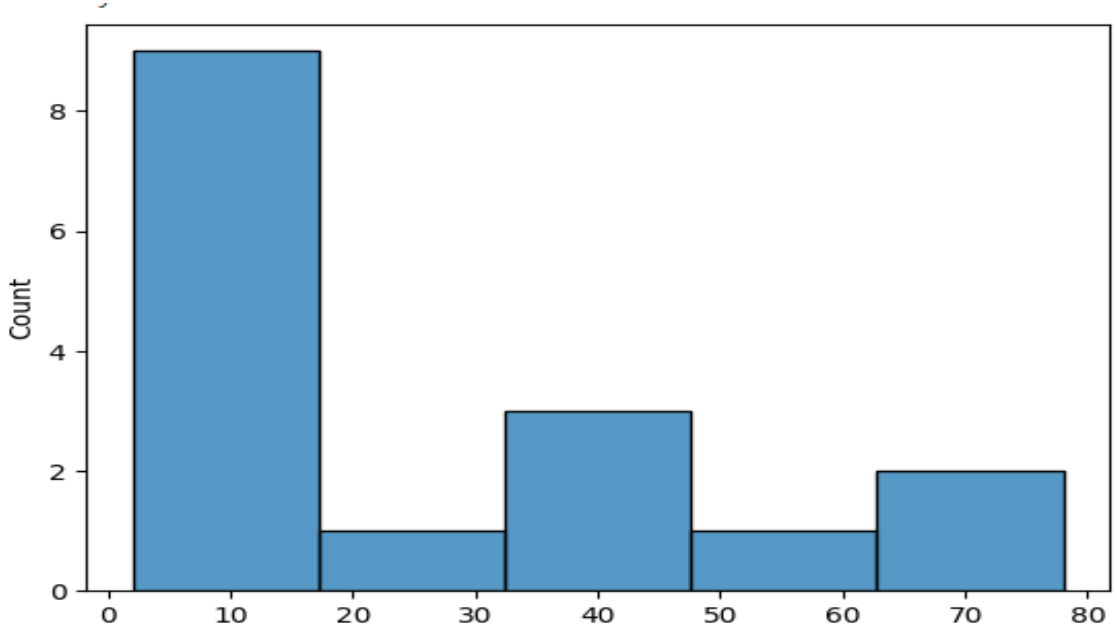
Distribution

Histogram: A histogram is a graphical representation of data points organized into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

data_points=[4,6,2,8,9,7,36,9,2,17,34,56,45,64,23,78]

In this diagram, we have divided the dataset into 5 logical ranges or bins.

For a histogram, x axis contains Continuous data points and y axis contains the frequency of each range

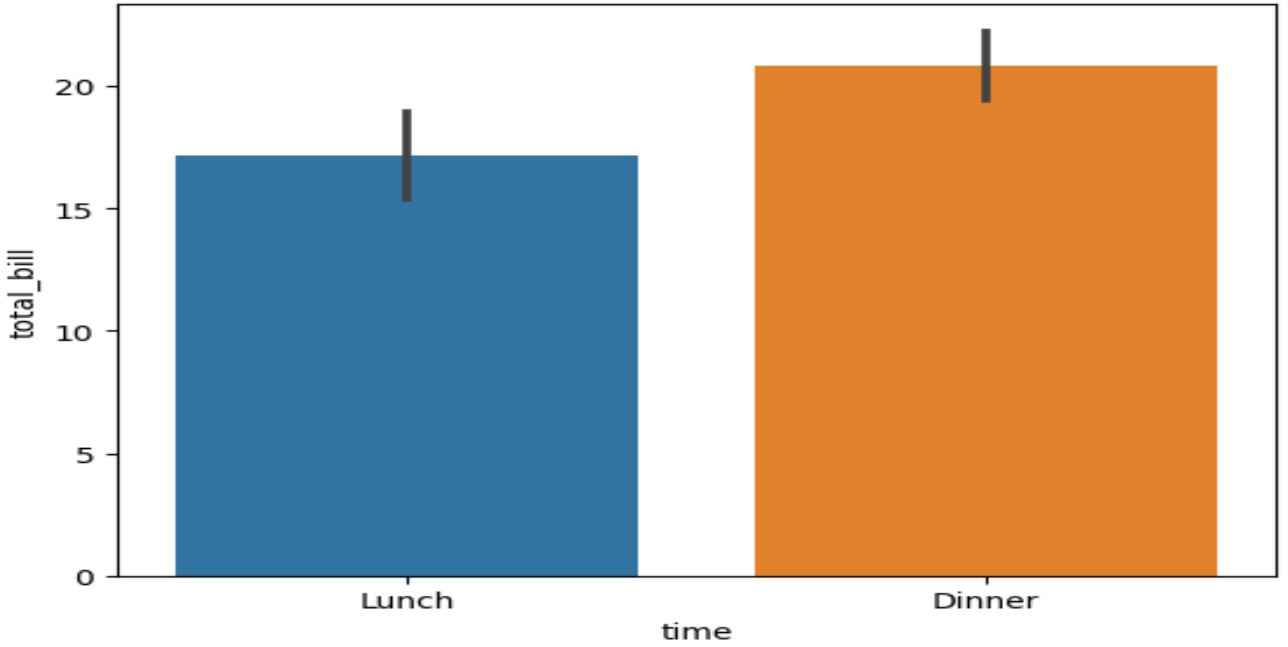


Distribution Contd.

Bar Graph: A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally.

A bar graph shows comparisons among discrete categories. One axis of the chart shows the specific categories being compared, and the other axis represents a measured value

The bar graph compares the total bills at a restaurant between lunch and dinner over a period of time.

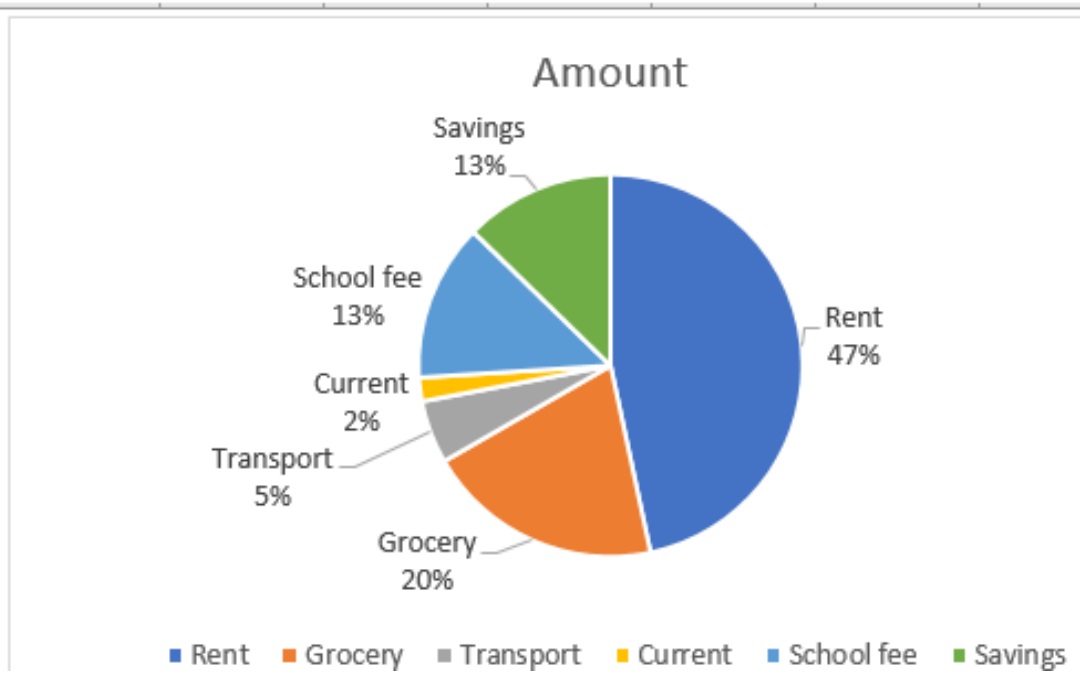


Distribution Contd.

Pie Chart: A pie chart is a circular statistical graphic which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice is proportional to the quantity it represents

Pie Chart Examples

	Expenses	Amount
2	Rent	7000
3	Grocery	3000
4	Transport	800
5	Current	300
6	School fee	2000
7	Savings	1900

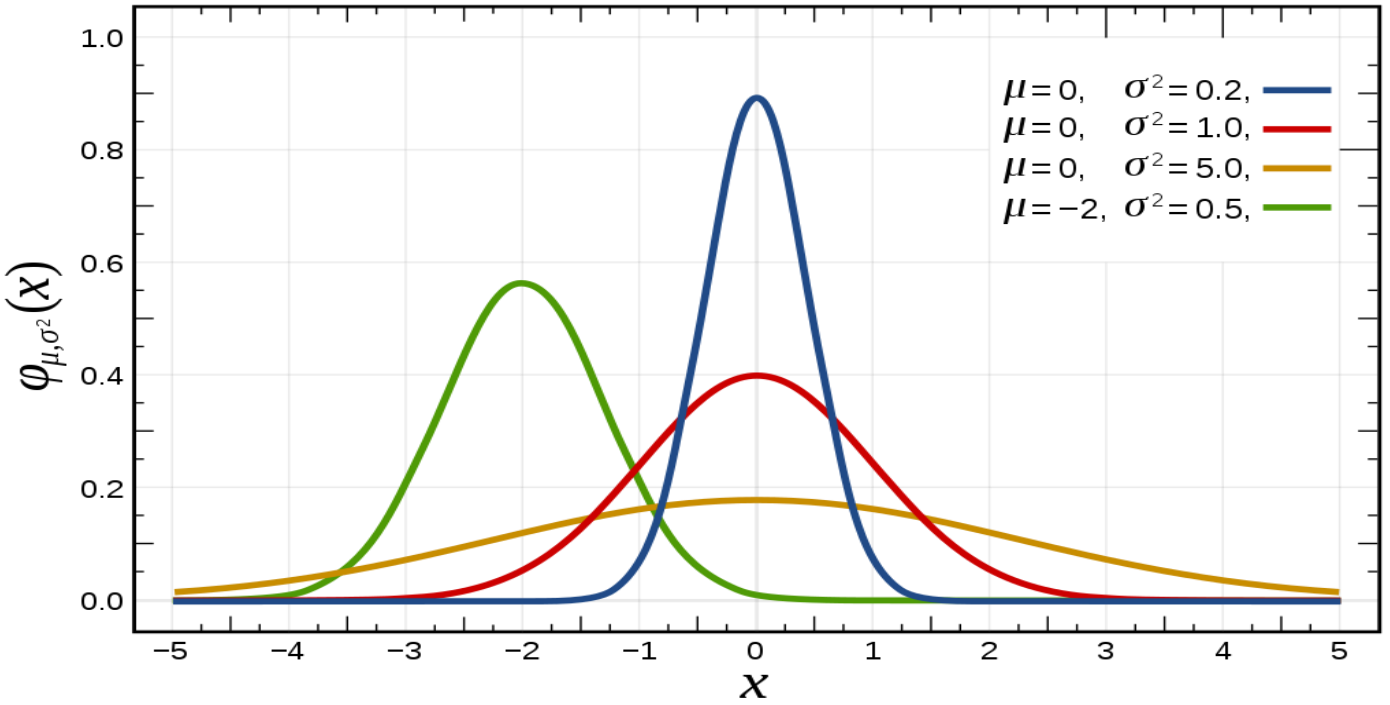


Distribution Contd.

Normal Distribution: Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graphical form, the normal distribution appears as a "bell curve".

PDF:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

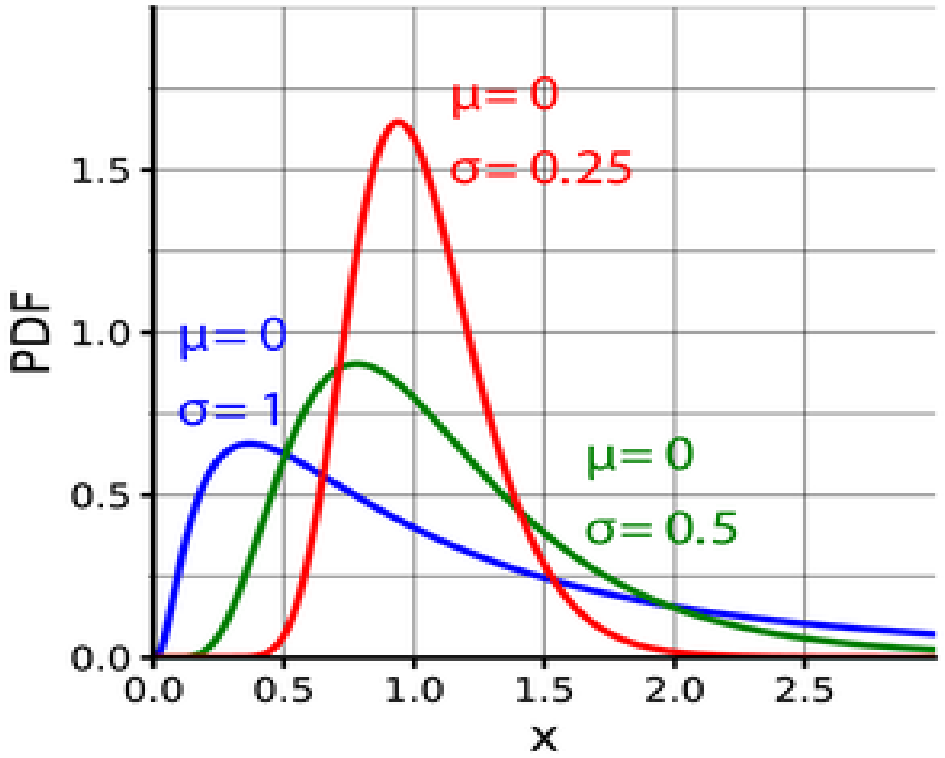


Distribution Contd.

Log Normal Distribution: In probability theory, a log-normal (or lognormal) distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. Thus, if the random variable X is log-normally distributed, then Y = ln(X) has a normal distribution.

PDF:

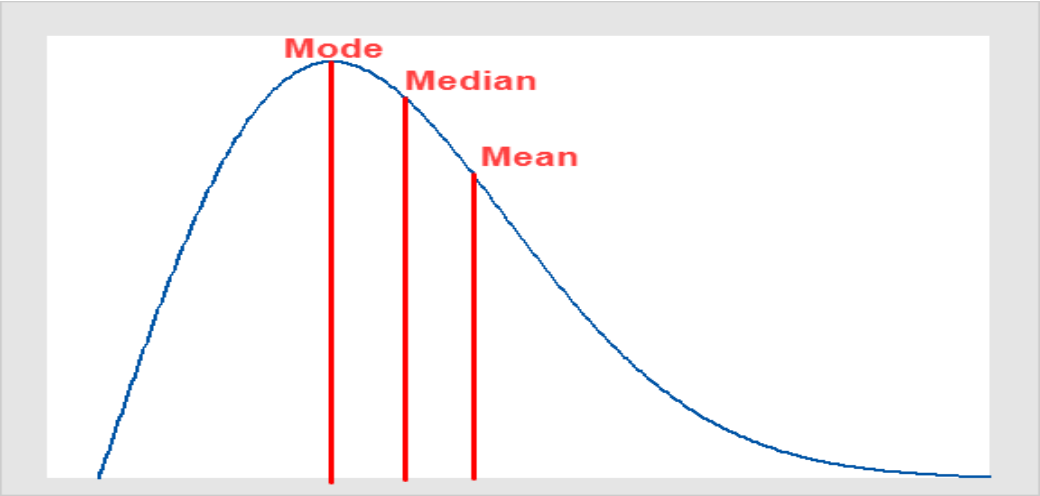
$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2}$$



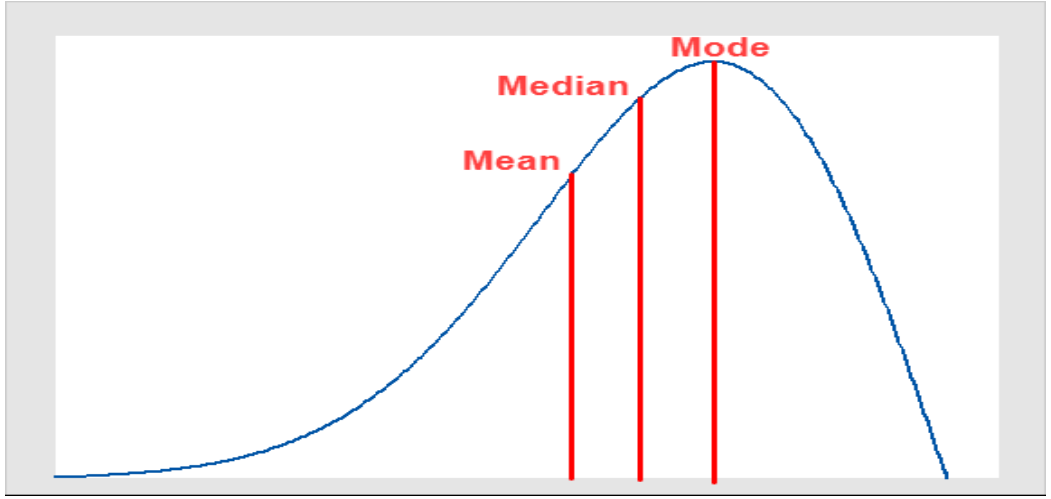
Distribution Contd.

Right Skewed Distribution (Positive skewness): Right skewed distributions occur when the long tail is on the right side of the distribution. Analysts also refer to them as positively skewed.

Left Skewed Distribution (Negative skewness): Left skewed distributions occur when the long tail is on the left side of the distribution. Statisticians also refer to them as negatively skewed



Right Skewed Distribution



Left Skewed Distribution

Distribution Contd.

Standard Normal Distribution: The standard normal distribution is one of the forms of the normal distribution. It occurs when a normal random variable has a mean equal to zero and a standard deviation equal to one. In other words, a normal distribution with a mean 0 and standard deviation of 1 is called the standard normal distribution. Also, the standard normal distribution is centered at zero, and the standard deviation gives the degree to which a given measurement deviates from the mean.

The random variable of a standard normal distribution is known as the standard score or a z-score. It is possible to transform every normal random variable X into a z score using the following formula:

$z = (X – \mu) / \sigma$

where X is a normal random variable, μ is the mean of X, and σ is the standard deviation of X

Let’s solve a problem statement using z score

Distribution Contd.

Problem 1: For some computers, the time period between charges of the battery is normally distributed with a mean of 50 hours and a standard deviation of 15 hours. Rohan has one of these computers and needs to know the probability that the time period will be between 50 and 70 hours?

Sol: $\mu = 50, \sigma = 15, X1 = 50, X2 = 70$

$z1 = (50\text{-}50)/15 = 0$

$z2 = (70\text{-}50)/15 = 1.33$

We will be using positive z table to find out the probabilities

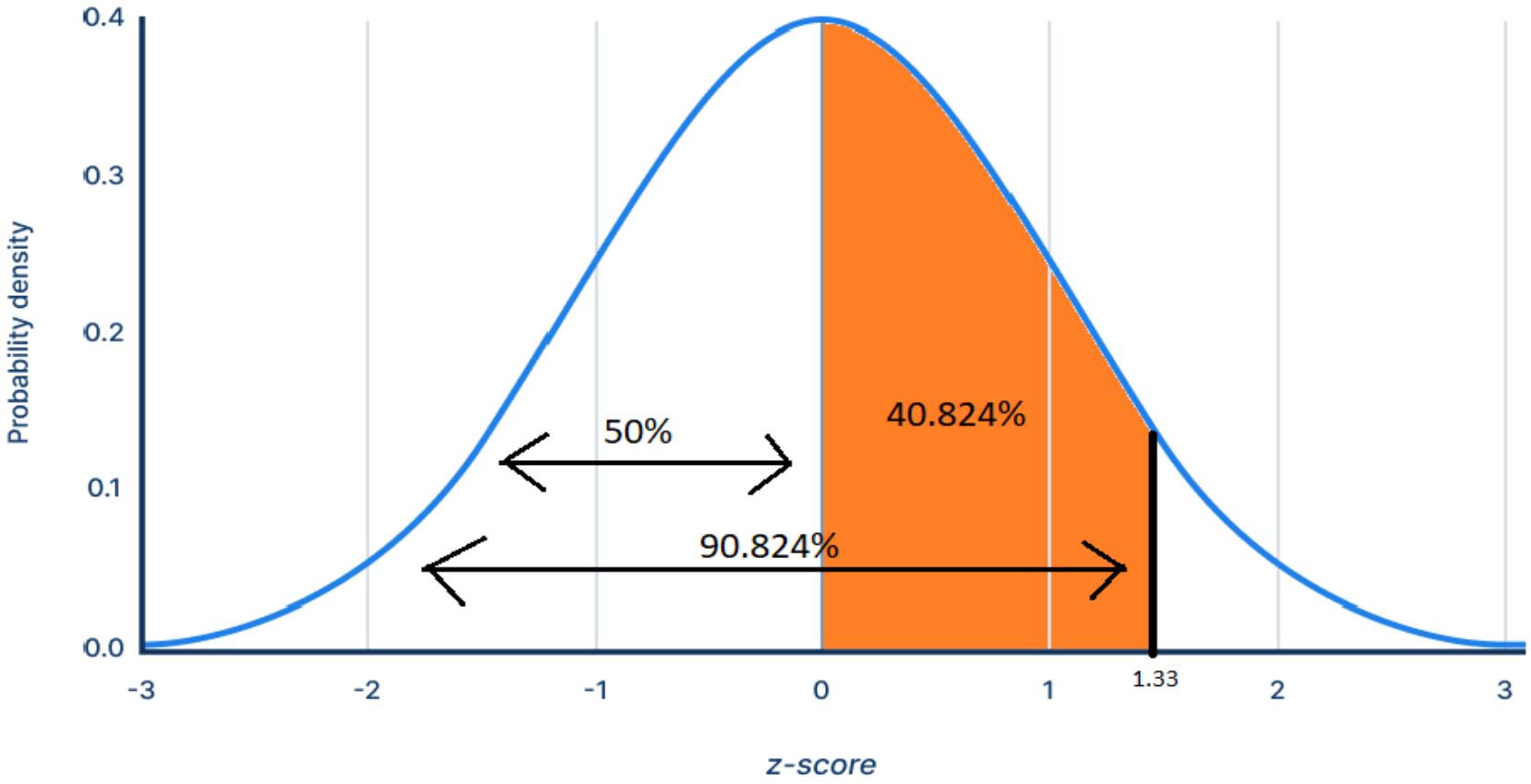
$P(x|z1<=0) = 0.5$

$P(x|z2<=1.33) = 0.90824$

$P(z1<x<z2) = (0.90824 - 0.5) = 0.40824$

So the probability that the time period will be between 50 and 70 hours is 0.40824 or 40.824%

Problem Statement 1 Visualization



Distribution Contd.

Problem 2: The speeds of cars are measured using a radar unit, on a motorway. The speeds are normally distributed with a mean of 90 km/hr and a standard deviation of 10 km/hr. What is the probability that a car selected at chance is moving at more than 100 km/hr?

Sol: $\mu = 90, \sigma = 10, X = 100$

$z = (100\text{-}90)/10 = 1$

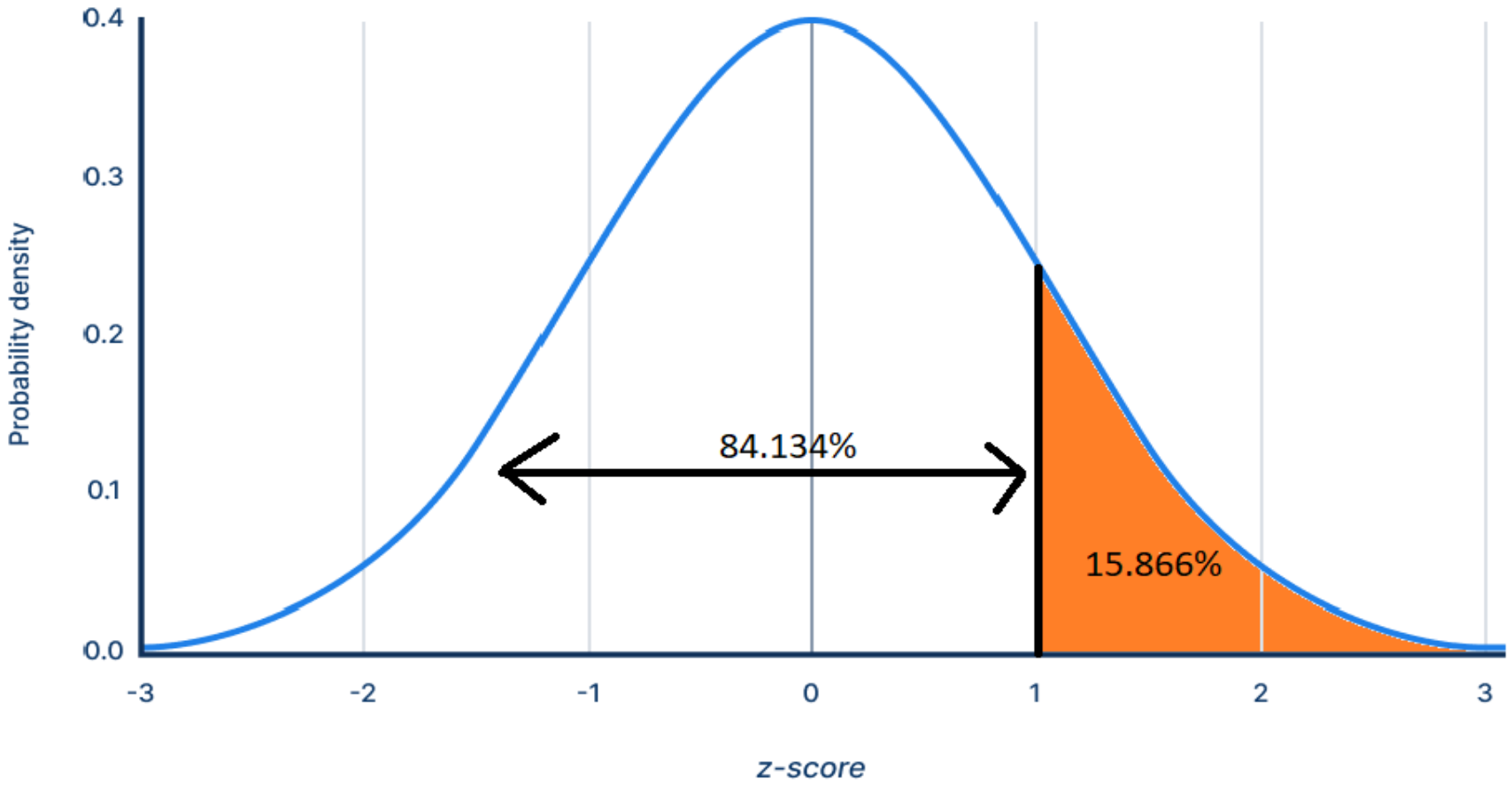
We will be using positive z table to find out the probability

$P(x|z<=1) = 0.84134$

$P(x|z>1) = (1\text{-}0.84134) = 0.15866$

So the probability that a car selected at chance is moving at more than 100 km/hr is 0.15866 or 15.866%

Problem Statement 2 Visualization



Inferential Statistics

While descriptive statistics summarize the characteristics of a data set, inferential statistics help you come to conclusions and make predictions based on your data.

When you have collected data from a sample, you can use inferential statistics to understand the larger population from which the sample is taken.

Types of Inferential statistics:

- Hypothesis testing
- Regression Analysis

Hypothesis testing

Hypothesis testing is a type of inferential statistics that is used to test assumptions and draw conclusions about the population from the available sample data. It involves setting up a null hypothesis and an alternative hypothesis followed by conducting a statistical test of significance. A conclusion is drawn based on the value of the test statistic, the critical value, and the confidence intervals. A hypothesis test can be left-tailed, right-tailed, and two-tailed.

Some important terminologies with respect to Hypothesis testing:

Null Hypothesis: The null-hypothesis is considered an accepted truth. It presumes that the sampled data and the population data have no difference. It is denoted by H_0

Alternative Hypothesis: An alternative hypothesis is an opposing theory to the null hypothesis. For example, if the null hypothesis predicts something to be true, the alternative hypothesis predicts it to be false. The alternative hypothesis often is the statement you test when attempting to disprove the null hypothesis. It is denoted by H_1 or H_a

Alpha level (Significance value): The significance level, also known as alpha or α, is a measure of the strength of the evidence that must be present in your sample before you will reject the null hypothesis and conclude that the effect is statistically significant

Confidence level: Also known as the confidence interval. This refers to how confident you can be that your conclusion is, in fact, correct. The confidence level is easy to calculate: the alpha and confidence levels always add up to one. ie:

$1 - \alpha$ = confidence level

Critical region: Set of all values which would cause us to reject the null hypothesis H_0. Also known as a rejection region.

Critical value(s): The value(s) which separate the critical region from the non-critical region. The critical values are determined independently of the sample statistics.

P-Value: A p-value is a crucial element of any hypothesis test results. It's a number between 0 and 1, and it gauges the probability that random fluctuations caused any data that might cause you to reject the null hypothesis. It's calculated by running test results through a statistical significance test. If the p-value is lower than your alpha level, then you reject the null hypothesis. If higher, then you do not reject the null hypothesis.

Steps to perform Hypothesis testing

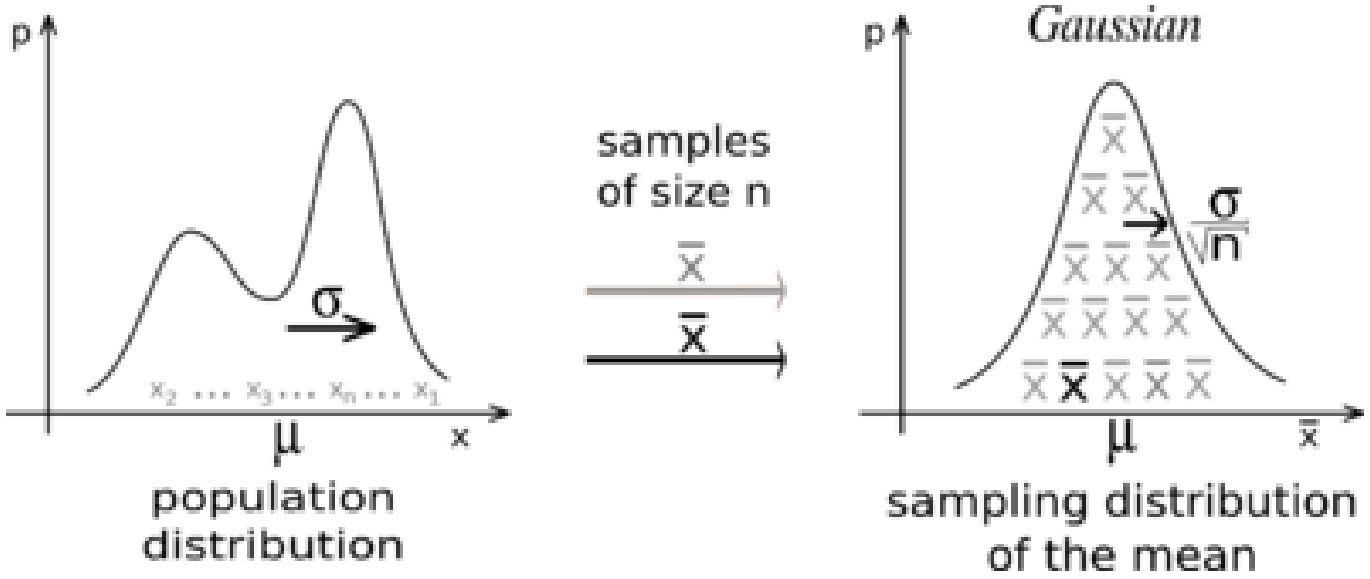
- Formulating the null hypothesis (H0) and alternative hypothesis (Ha): The null hypothesis is the statement of no effect or no difference in the population, while the alternative hypothesis is the statement that contradicts the null hypothesis and represents the possibility of an effect or difference.
- Choosing a significance level (α): The significance level, denoted as α, is the predetermined threshold that determines the level of evidence required to reject the null hypothesis.
- Selecting an appropriate statistical test: The choice of a statistical test depends on the nature of the data and the research question. There are various tests available, such as t-tests, chi-square tests, ANOVA, correlation analysis, and regression analysis, among others.
- Collecting and analyzing data: Data is collected from a sample, and the statistical test is applied to the sample data. The test generates a test statistic, which measures the degree of evidence against the null hypothesis. The test statistic follows a specific probability distribution under the assumption of the null hypothesis.

- Determining the critical region and p-value: The critical region is the range of values of the test statistic that leads to the rejection of the null hypothesis. The p-value is the probability of obtaining a test statistic as extreme as or more extreme than the one observed, assuming the null hypothesis is true. If the p-value is smaller than the chosen significance level (α), the null hypothesis is rejected.
- Interpreting the results: Based on the analysis, a decision is made regarding the null hypothesis. If the null hypothesis is rejected, it suggests that there is sufficient evidence to support the alternative hypothesis. If the null hypothesis is not rejected, it indicates that there is not enough evidence to support the alternative hypothesis.

Central Limit Theorem

- The central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.

In order to satisfy the Central Limit theorem, the minimum sample size should be 30.



- The mean of the sampling distribution is the mean of the population $\mu_{\bar{x}} = \mu$
- The standard deviation of the sampling distribution is the standard deviation of the population divided by the square root of the sample size. $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

We can describe the sampling distribution of the mean using this notation:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Where:

- X̄ is the sampling distribution of the sample means
- ~ means “follows the distribution”
- N is the normal distribution
- μ is the mean of the population
- σ is the standard deviation of the population
- n is the sample size

Z-test

- A z test is conducted on a population that follows a normal distribution with independent data points and has a sample size that is greater than or equal to 30. It is used to check whether the means of two populations are equal to each other when the population variance is known. The null hypothesis of a z test can be rejected if the z test statistic is statistically significant when compared with the critical value.

Left Tailed Test:

- Null Hypothesis: $H0 : \mu=\mu0$
- Alternate Hypothesis: $H1 : \mu<\mu0$
- Decision Criteria: If the z statistic < z critical value then reject the null hypothesis.

Let’s take an example of left tailed Z test.

Example 1: An online medicine shop claims that the mean delivery time for medicines is less than 120 minutes with a standard deviation of 30 minutes. Is there enough evidence to support this claim at a 0.05 significance level if 49 orders were examined with a mean of 100 minutes?

Solution: As the sample size is 49 and population standard deviation is known, this is an example of a left-tailed one-sample z test.

H0 : $\mu=120$

H1 : $\mu<120$

From the z table, the critical value = -1.64 (at $\alpha=0.05$). A negative sign is used as this is a left tailed test.

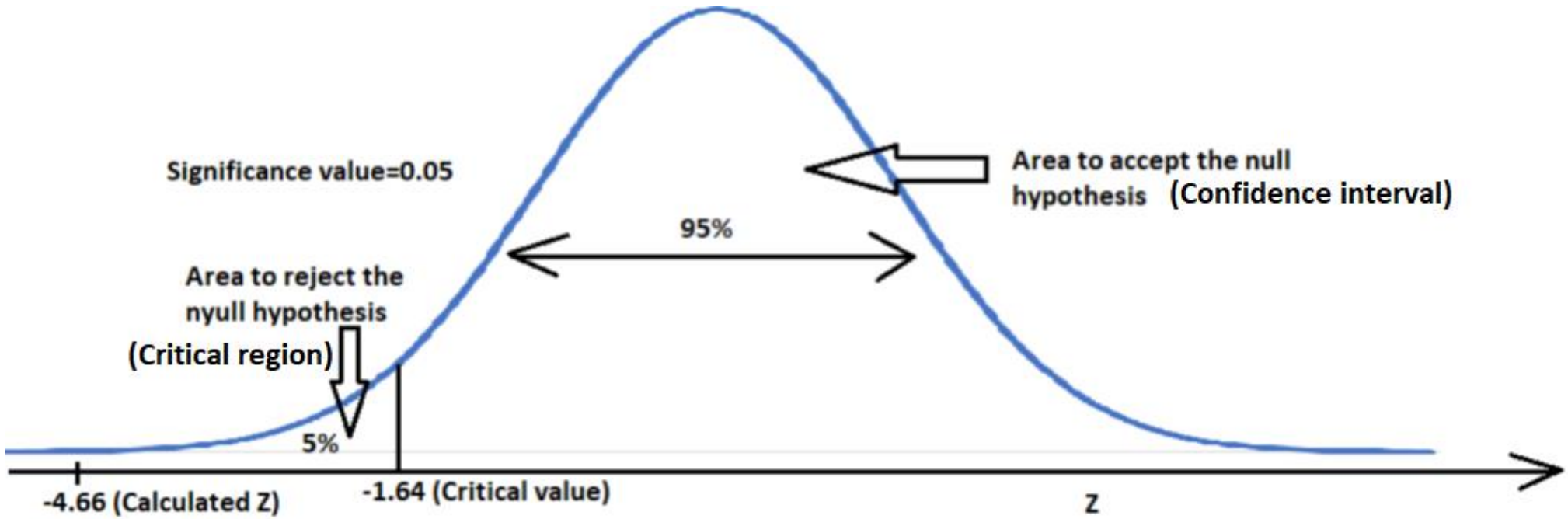
$$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

$\bar{x}$ = 100, μ= 120, n = 49, σ= 30

z = -4.66

As -4.66 < -1.645 thus, the null hypothesis is rejected and it is concluded that there is enough evidence to support the medicine shop's claim.

Answer: Reject the null hypothesis


Significance value=0.05
Area to accept the null hypothesis (Confidence interval)
95%
Area to reject the nyull hypothesis
(Critical region)
5%
-4.66 (Calculated Z)
-1.64 (Critical value)
Z

Right Tailed Test:

- Null Hypothesis: H0: $\mu=\mu0$
- Alternate Hypothesis: H1 : $\mu>\mu0$
- Decision Criteria: If the z statistic > z critical value then reject the null hypothesis.

Example 2: A teacher claims that the mean score of students in his class is greater than 82 with a standard deviation of 20. If a sample of 81 students was selected with a mean score of 90 then check if there is enough evidence to support this claim at a 0.05 significance level.

Solution: As the sample size is 81 and population standard deviation is known, this is an example of a right-tailed one-sample z test.

H0: $\mu=82$

H1: $\mu>82$

From the z table, the critical value = 1.64 (at $\alpha=0.05$)

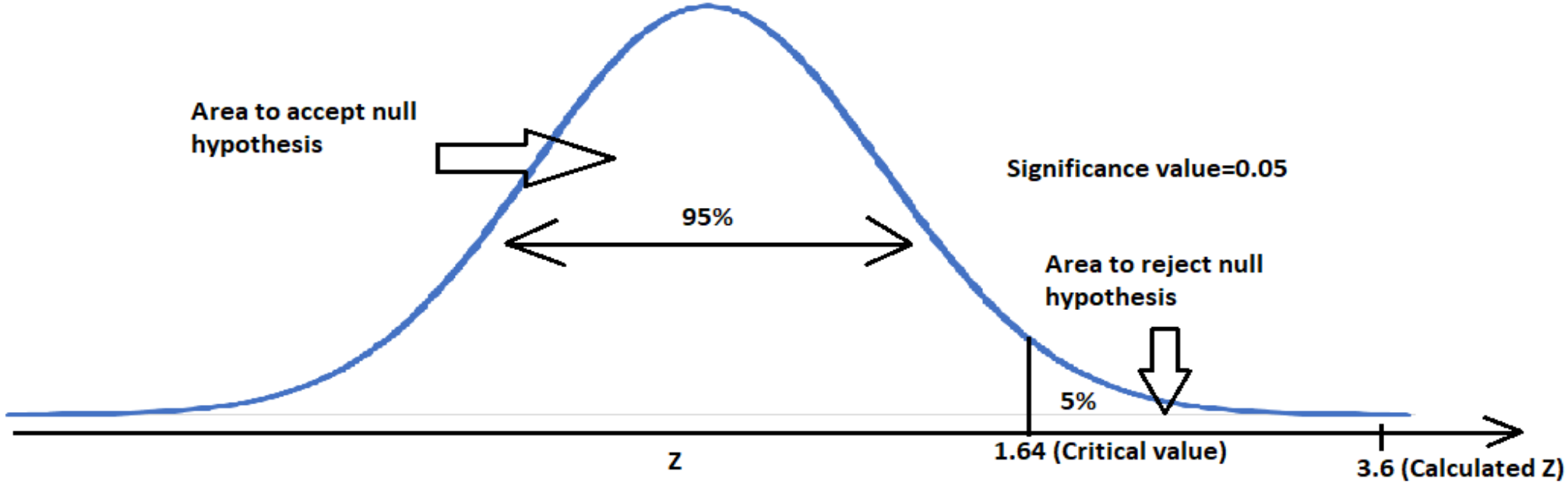
$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$\bar{x}$ = 90, μ= 82, n = 81, σ= 20

z = 3.6

As 3.6 > 1.64 thus, the null hypothesis is rejected and it is concluded that there is enough evidence to support the teacher's claim.

Answer: Reject the null hypothesis



Two Tailed Test:

- Null Hypothesis: H0 : $\mu=\mu0$
- Alternate Hypothesis: H1: $\mu\neq\mu0$
- Decision Criteria: If the z statistic > z_{right} critical value or the z statistic < z_{left} then reject the null hypothesis.

Example 3: The average height of students in a batch is 100 cm and the standard deviation is 15. However, Tedd believes that this has changed, so he decides to test the height of 75 random students in the batch. The average height of the sample comes out to be 105. Is there enough evidence to suggest that the average height has changed?

Solution: This is an example of two tailed one sample Z test

H0: $\mu=82$

H1: $\mu\neq82$

As significance value is not provided in this question, we will assume $\alpha=0.05$

This is a two tailed test, so the significance value will be divided in two equal parts (0.025) between the left tail and right tail respectively.

From Z table, we get $Z_{leftcritical} = -1.96$, $Z_{rightcritical} = 1.96$

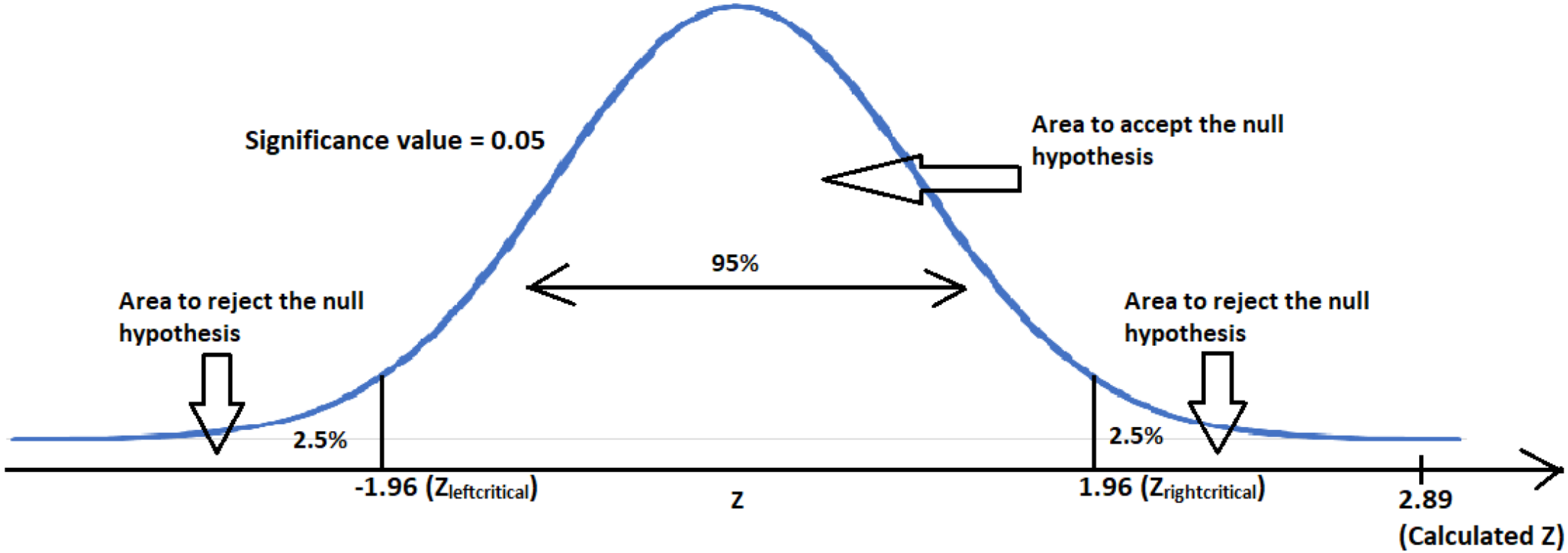
$$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

$\bar{x} = 105$, $\mu = 100$, $n = 75$, $\sigma = 15$

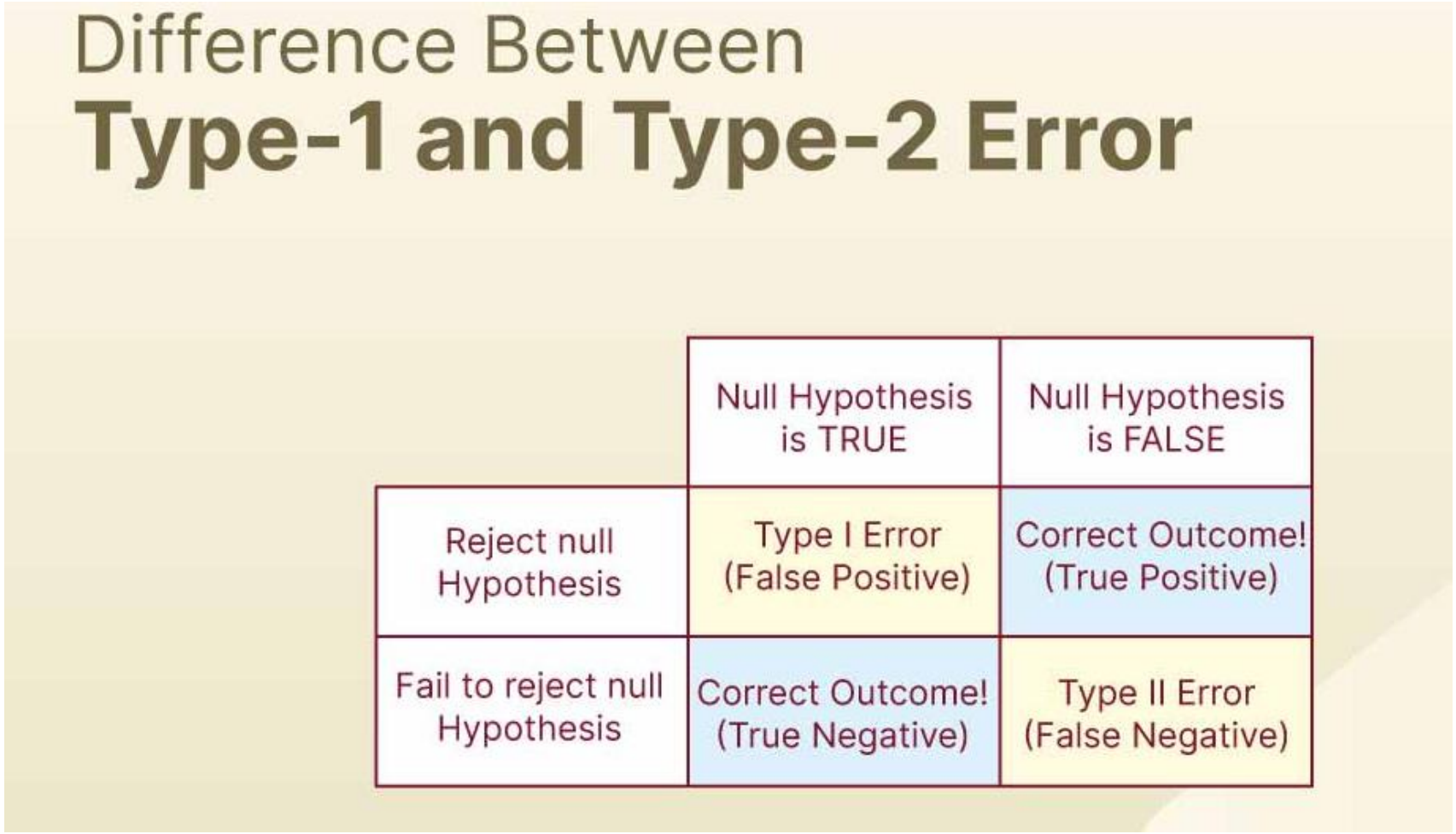
$Z = 2.89$

As Z does not lie between -1.96 to 1.96, we can reject the null hypothesis and it is concluded that there is enough evidence to support Tedd's claim

Answer: Reject the null hypothesis


Significance value = 0.05
Area to accept the null hypothesis
95%
Area to reject the null hypothesis
Area to reject the null hypothesis
2.5%
2.5%
-1.96 (Zleftcritical)
Z
1.96 (Zrightcritical)
2.89
(Calculated Z)

Type-I and Type-II error



T-test

A t-test is an inferential statistic used to determine if there is a significant difference between the means of two groups and how they are related. T-tests are used when the data sets follow a normal distribution and have unknown variances.

Types of T-test:

- **One-sample t-test** — compare the mean of one group against the specified mean generated from a population. For example, a manufacturer of mobile phones promises that one of their models has a battery that supports about 25 hours of video playback on average. To find out if the manufacturer is right, a researcher can sample 15 phones, measure the battery life and get an average of 23 hours. Then, he can use a t-test to determine whether this difference is received not just by chance.

- **Paired sample t-test** — compares the means of two measurements taken from the same individuals, objects, or related units. For instance, students passed an additional course for math and it would be interesting to find whether their results became better after course completion. It is possible to take a sample from the same group and use the paired t-test.
- **An Independent two-sample t-test** —is used to analyze the mean comparison of two independent groups. Like two groups of students.

We will discuss about one sample t-test which is the basic one

t-score formula (one sample t-test):

$$t=\frac{(\bar{X}-\mu)}{S/\sqrt{n}}$$

Where, $\bar{X}$ is the sample mean, μ is the hypothesized population mean, S is the standard deviation of the sample and n is the number of sample observations

When working with T-test, T-distribution is used in place of the normal distribution. The t-distribution is a family of curves that are symmetrical about the mean, and have increasing variability as the degrees of freedom increase. The t-test statistic (T) follows a t-distribution with n – 1 degrees of freedom, where n is the number of observations in the sample.

One-sample T-test: Example

Suppose a claim is made that the average number of days a person spends on vacation is more than or equal to 5 days (hypothesized population mean) based on a sample of 16 people whose mean came out to be 9 days, and standard deviation as 3. Is there enough evidence to support the claim?

Solution:

Null hypothesis, H0: There is no difference between the sample mean and the population mean; What has occurred with a sample is just an instance of chance occurrence.

Alternate hypothesis, Ha: There is a significant difference between the sample mean and the population mean. We will use one-sample t-test to test this hypothesis. A right-tailed test will be performed.

As we know $\bar{X}=9$, $\mu=5$, $S=3$, $n=16$

$$t=\frac{(\bar{X}-\mu)}{S/\sqrt{n}} = 5.33$$

We can assume significance value as 0.05.

Degree of freedom = n-1 = 16-1 = 15

With the help of significance value and the degree of freedom, we can find out the $t_{critical}$ from t-table.

$t_{critical}$ = 1.753

Since the calculated t-value of 5.33 is much larger than the critical value of 1.753, the null hypothesis can be rejected. Thus, there is a statistically significant difference between sample mean and the population mean.

Another way to test is to calculate the p-value for getting the T-statistics of 5.33. we can use any P-value calculator to calculate p-value for a given T-value, degrees of freedom and the types of tail-test (one-tailed or two-tailed test). For a T-statistics of 5.33, the p-value came out to be 0.000042. This means that there is a probability of only 0.000042 to get this kind of sample given the null hypothesis holds good. As this value is less than 0.05, one can reject the null hypothesis given the evidence of current sample.

If P-value <= Significance value => Reject the null hypothesis

If P-value > Significance value => Fail to reject the null hypothesis

Chi-Square test

A Pearson's chi-square test is a statistical test for categorical data. It is used to determine whether your data are significantly different from what you expected

If we want to test a hypothesis about the distribution of a **categorical variable** you'll need to use a chi-square test or another nonparametric test. Categorical variables can be nominal or ordinal and represent groupings such as species or nationalities. Because they can only have a few specific values, they can't have a normal distribution.

There are two types of Pearson's chi-square tests:

- The **chi-square goodness of fit test** is used to test whether the frequency distribution of a categorical variable is different from your expectations.
- The **chi-square test of independence** is used to test whether two categorical variables are related to each other.

Here we will discuss the first type of chi-square test which is goodness of fit test

.

The chi-square formula

Both of Pearson's chi-square tests use the same formula to calculate the test statistic, chi-square (X^2):

$$X^2 = \sum \frac{(O - E)^2}{E}$$

Where:

- X^2 is the chi-square test statistic
- Σ is the summation operator (it means "take the sum of")
- O is the observed frequency
- E is the expected frequency

The larger the difference between the observations and the expectations ($O - E$ in the equation), the bigger the chi-square will be. To decide whether the difference is big enough to be statistically significant, you compare the chi-square value to a critical value.

Example:

After weeks of hard work, your dog food experiment is complete and you compile your data in a table:

Observed and expected frequencies of dogs' flavor choices

Flavor	Observed	Expected
Garlic Blast	22	25
Blueberry Delight	30	25
Minty Munch	23	25

Would you conclude that the frequencies of dog's flavor choices are in different proportions? (Significance level = 0.05)

Solution:

Null hypothesis (H0): The dog population chooses the three flavors in equal proportions (p1 = p2 = p3).

Alternative hypothesis (H_a): The dog population does not choose the three flavors in equal proportions.

Degree of freedom (df) = number of groups – 1 = 3 – 1 = 2

For a test of significance at α = .05 and df = 2, the X^2 critical value is 5.99.

Let's calculate the chi-square critical value

$$X^2 = \sum \frac{(O-E)^2}{E}$$

$X^2 = \frac{(22-25)^2}{25} + \frac{(30-25)^2}{25} + \frac{(23-25)^2}{25} = 1.52$

Also, p-value = 0.4677 (we can find this with any p-value calculator available in the internet)

The X^2 value is less than the critical value (1.52<5.99). Therefore, we should not reject the null hypothesis that the dog population chooses the three flavors in equal proportions. There is no significant difference between the observed and expected flavor choice distribution (p >0 .05). This suggests that the dog food flavors are equally popular in the dog population.

ANOVA (Analysis of Variance)

An ANOVA test is a statistical test used to determine if there is a statistically significant difference between two or more categorical groups by testing for differences of means using a variance.

Assumptions Of ANOVA

- The assumptions of the ANOVA test are the same as the general assumptions for any parametric test:
- An ANOVA can only be conducted if there is **no relationship between the subjects** in each sample. This means that subjects in the first group cannot also be in the second group (e.g., independent samples/between groups).
- The different groups/levels must have **equal sample sizes**.
- An ANOVA can only be conducted if the dependent variable is **normally distributed** so that the middle scores are the most frequent and the extreme scores are the least frequent.
- Population variances must be equal (i.e., homoscedastic). Homogeneity of variance means that the deviation of scores (measured by the range or standard deviation, for example) is similar between populations.

Types Of ANOVA Tests:

- One way ANOVA
- Two way ANOVA

One way ANOVA:

A one-way ANOVA (analysis of variance) has one categorical independent variable (also known as a factor) and a normally distributed continuous (i.e., interval or ratio level) dependent variable.

The independent variable divides cases into two or more mutually exclusive levels, categories, or groups.

The one-way ANOVA test for differences in the means of the dependent variable is broken down by the levels of the independent variable.

An example of a one-way ANOVA includes testing a therapeutic intervention (CBT, medication, placebo) on the incidence of depression in a clinical sample.

Note: Both the One-Way ANOVA and the Independent Samples t-Test can compare the means for two groups. However, only the One-Way ANOVA can compare the means across three or more groups.

Two-way (factorial) ANOVA

A two-way ANOVA (analysis of variance) has two or more categorical independent variables (also known as a factor) and a normally distributed continuous (i.e., interval or ratio level) dependent variable.

The independent variables divide cases into two or more mutually exclusive levels, categories, or groups. A two-way ANOVA is also called a factorial ANOVA.

An example of factorial ANOVAs include testing the effects of social contact (high, medium, low), job status (employed, self-employed, unemployed, retired), and family history (no family history, some family history) on the incidence of depression in a population.

We will take an example of One way ANOVA to understand the concept

Example:

Suppose we want to know whether or not three different exam prep programs lead to different mean scores on a certain exam. To test this, we recruit 30 students to participate in a study and split them into three groups.

The students in each group are randomly assigned to use one of the three exam prep programs for the next three weeks to prepare for an exam. At the end of the three weeks, all of the students take the same exam.

The exam scores for each group are shown below:

Group 1	Group 2	Group 3
85	91	79
86	92	78
88	93	88
75	85	94
78	87	92
94	84	85
98	82	83
79	88	85
71	95	82
80	96	81

Perform an ANOVA test with this dataset.

Solution:

H_0 (null hypothesis): $\mu_1 = \mu_2 = \mu_3$ (The exam prep programs does not make any difference)

H_1 (alternative hypothesis): at least one exam prep program makes the difference from the rest

Group 1	Group 2	Group 3
85	91	79
86	92	78
88	93	88
75	85	94
78	87	92
94	84	85
98	82	83
79	88	85
71	95	82
80	96	81
$\overline{X1} = 83.4$	$\overline{X2} = 89.3$	$\overline{X3} = 84.7$

$\overline{\overline{X}} = 85.8$

Error sum of square (SSE):

Group 1	Group 2	Group 3
85	91	79
86	92	78
88	93	88
75	85	94
78	87	92
94	84	85
98	82	83
79	88	85
71	95	82
80	96	81
$\overline{X1} = 83.4$	$\overline{X2} = 89.3$	$\overline{X3} = 84.7$

$\overline{\overline{X}} = 85.8$

Group 1

Group 2

Group 3

$SSE = (85-83.4)^2 + (86-83.4)^2 + (88-83.4)^2 + (75-83.4)^2 + (78-83.4)^2 + (94-83.4)^2 + (98-83.4)^2 + (79-83.4)^2 + (71-83.4)^2 + (80-83.4)^2 + (91-89.3)^2 + (92-89.3)^2 + (93-89.3)^2 + (85-89.3)^2 + (87-89.3)^2 + (84-89.3)^2 + (82-89.3)^2 + (88-89.3)^2 + (95-89.3)^2 + (96-89.3)^2 + (79-84.7)^2 + (78-84.7)^2 + (88-84.7)^2 + (94-84.7)^2 + (92-84.7)^2 + (85-84.7)^2 + (83-84.7)^2 + (85-84.7)^2 + (82-84.7)^2 + (81-84.7)^2$

$SSE = 640.4 + 208.1 + 252.1$

$SSE = 1100.6$

Sum of Square between the groups (SSB):

Group 1	Group 2	Group 3
85	91	79
86	92	78
88	93	88
75	85	94
78	87	92
94	84	85
98	82	83
79	88	85
71	95	82
80	96	81
$\overline{X1} = 83.4$	$\overline{X2} = 89.3$	$\overline{X3} = 84.7$

$\overline{\overline{X}} = 85.8$

$SSB = (83.4-85.8)^2 + (83.4-85.8)^2 + (83.4-85.8)^2 + (83.4-85.8)^2 + (83.4-85.8)^2 + (83.4-85.8)^2 + (83.4-85.8)^2 + (83.4-85.8)^2 + (83.4-85.8)^2 + (83.4-85.8)^2 + (89.3-85.8)^2 + (89.3-85.8)^2 + (89.3-85.8)^2 + (89.3-85.8)^2 + (89.3-85.8)^2 + (89.3-85.8)^2 + (89.3-85.8)^2 + (89.3-85.8)^2 + (89.3-85.8)^2 + (89.3-85.8)^2 + (84.7-85.8)^2 + (84.7-85.8)^2 + (84.7-85.8)^2 + (84.7-85.8)^2 + (84.7-85.8)^2 + (84.7-85.8)^2 + (84.7-85.8)^2 + (84.7-85.8)^2 + (84.7-85.8)^2 + (84.7-85.8)^2$

$SSB = 57.6 + 122.5 + 12.1$

$SSB = 192.2$

Total Sum of square (SST):

SST = SSE + SSB = 1100.6 + 192.2 = 1292.8

Error degree of freedom (df_e) = (n-k), where n is the total number of data points and k is the number of groups

df_e = (30 -3) = 27

Degree of freedom between the goups (df_b) = (k-1), where k is the number of groups

df_b = (3-1) = 2

Total degree of freedom (df_t) = (n-1) = df_e + df_b = 29

Error mean square (MSE):

MSE = $\frac{SSE}{df_e} = \frac{1100.6}{27}$ = 40.76

Between the group mean square (MSB):

MSB = $\frac{SSB}{\mathrm{df}_b} = \frac{192.2}{2} = 96.1$

ANOVA F score formula:

F score = $\frac{MSB}{MSE} = \frac{96.1}{40.76} = 2.358$

The corresponding p-value of the F-score is 0.1138

Let's assume significance value as 0.05

Since p-value > 0.05, we fail to reject the null hypothesis.

This means we don't have sufficient evidence to say that there is a statistically significant difference between the mean exam scores of the three groups.

Note that the ANOVA alone does not tell us specifically which means were different from one another. To determine that, we would need to follow up with multiple comparisons (or post-hoc) tests.

F-test

F test is a statistical test that is used in hypothesis testing to check whether the **variances of two populations or two samples are equal or not**. In an f test, the data follows an f distribution. This test uses the f statistic to compare two variances by dividing them. An f test can either be one-tailed or two-tailed depending upon the parameters of the problem.

This test is also called variance ratio test.

Formula:

- F statistic for large samples: $F = \frac{\sigma_1^2}{\sigma_2^2}$ where σ_1 is the variance of the first population and σ_2 is the variance of the second population.
- F statistic for small samples: $F = \frac{S_1^2}{S_2^2}$, where S_1 is the variance of the first sample and S_2 is the variance of the second sample.

The selection criteria for the σ_1^2 and σ_2^2 for an f statistic is given below:

- For a right-tailed and a two-tailed f test, the variance with the greater value will be in the numerator. Thus, the sample corresponding to σ_1^2 will become the first sample. The smaller value variance will be the denominator and belongs to the second sample.
- For a left-tailed test, the smallest variance becomes the numerator (sample 1) and the highest variance goes in the denominator (sample 2).

The steps to find the f test critical value at a specific alpha level (or significance level), α, are as follows:

- Find the degrees of freedom of the first sample. This is done by subtracting 1 from the first sample size. Thus, $x = n_1 - 1$.
- Determine the degrees of freedom of the second sample by subtracting 1 from the sample size. This given $y = n_2 - 1$
- If it is a right-tailed test then α is the significance level. For a left-tailed test 1 - α is the alpha level. However, if it is a two-tailed test then the significance level is given by α / 2.
- The F table is used to find the critical value at the required alpha level.
- The intersection of the x column and the y row in the f table will give the f test critical value.

Let's take an example to understand the F-test

Example: The bank has a head office in Delhi and a branch in Mumbai. There are long customer queues at one office, while customer queues are short at the other. The Operations Manager of the bank wonders if the customers at one branch are more variable than the number of customers at another. He carries out a research study of customers.

The variance of Delhi head office customers is 31, and that for the Mumbai branch is 20. The sample size for the Delhi head office is 11, and that for the Mumbai branch is 21. Carry out a two-tailed F-test with a level of significance of 10%.

Solution:

- **Step 1:** Null Hypothesis H_0: $\sigma_1^2 = \sigma_2^2$
- Alternate Hypothesis H_a: $\sigma_1^2 \neq \sigma_2^2$
- **Step 2:** F statistic = F Value = σ_1^2 / σ_2^2 = 31/20 = **1.55**
- **Step 3:** $df_1 = n_1 - 1$ = 11-1 = 10
- $df_2 = n_2 - 1$ = 21-1 = 20
- **Step 4:** Since it is a two-tailed test, alpha level = 0.10/2 = 0.05. The F value from the F Table with degrees of freedom as 10 and 20 is 2.348.
- **Step 5:** Since the F statistic (1.55) is lesser than the table value obtained (2.348), we cannot reject the null hypothesis.

Comparison between Z, T, Chi-Square, F test

Test statistic	Associated test	Sample size	Information given	Distribution	Test question	Parametric ?
z-score	z-test	Two populations or large samples ($n > 30$)	• Standard deviation of the population (this will be given as σ) • Population mean or proportion	Normal	Do these two populations differ?	YES
t-statistic	t-test	Two small samples ($n < 30$)	• Standard deviation of the sample (this will be given as s) • Sample mean	Normal	Do these two samples differ?	YES
f-statistic	ANOVA	Three or more samples	• Group sizes • Group means • Group standard deviations	Normal	Do any of these three or more samples differ from each other?	YES
chi-squared	chi-squared test	Two samples	• Number of observations for each categorical variable	Any	Are these two categorical variables independent?	NO

Regression Analysis

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.

Types of regression analysis:

1. Simple Linear Regression
2. Multiple Linear Regression
3. Nonlinear Regression

Simple Linear Regression:

Linear regression is used to predict the relationship between two variables by applying a linear equation to observed data. There are two types of variable, one variable is called an independent variable, and the other is a dependent variable. Linear regression is commonly used for predictive analysis.

Examples of Simple Linear Regression

The weight of the person is linearly related to their height. So, this shows a linear relationship between the height and weight of the person. According to this, as we increase the height, the weight of the person will also increase

Simple Linear Regression Equation

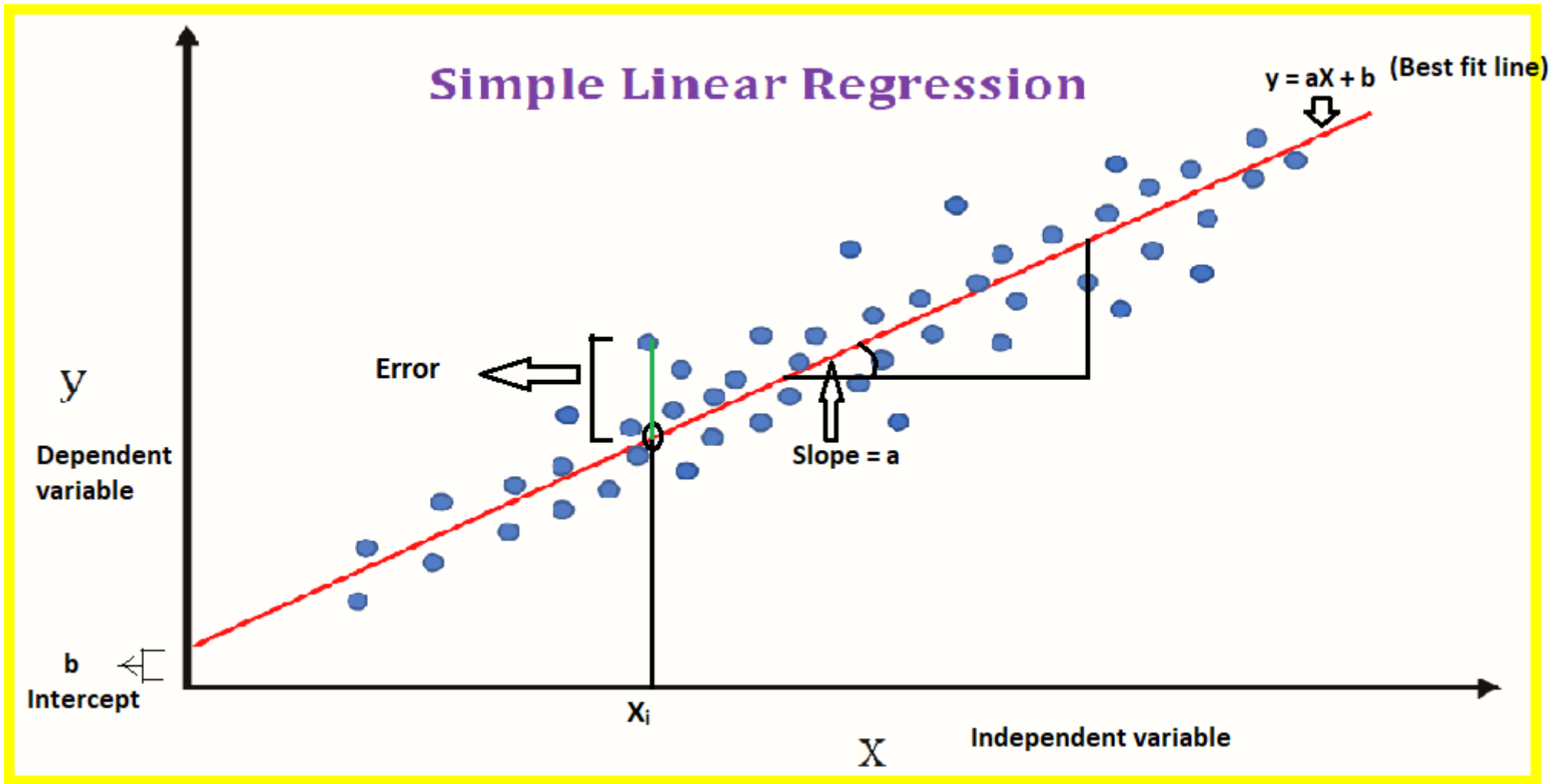
Linear Regression Equation is given below:

$Y=a+bX$

where X is the independent variable and it is plotted along the x-axis

Y is the dependent variable and it is plotted along the y-axis

Here, the slope of the line is b, and a is the intercept (the value of y when x = 0).


Simple Linear Regression
y = aX + b
(Best fit line)
Error
Slope = a
y
Dependent variable
b
Intercept
X_i
X
Independent variable

Multiple Linear Regression:

Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable.

Examples of Multiple Linear Regression:

Amount of rainfall depends on several factors, such as temperature, humidity, cloud cover etc.

Multiple Linear Regression equation:

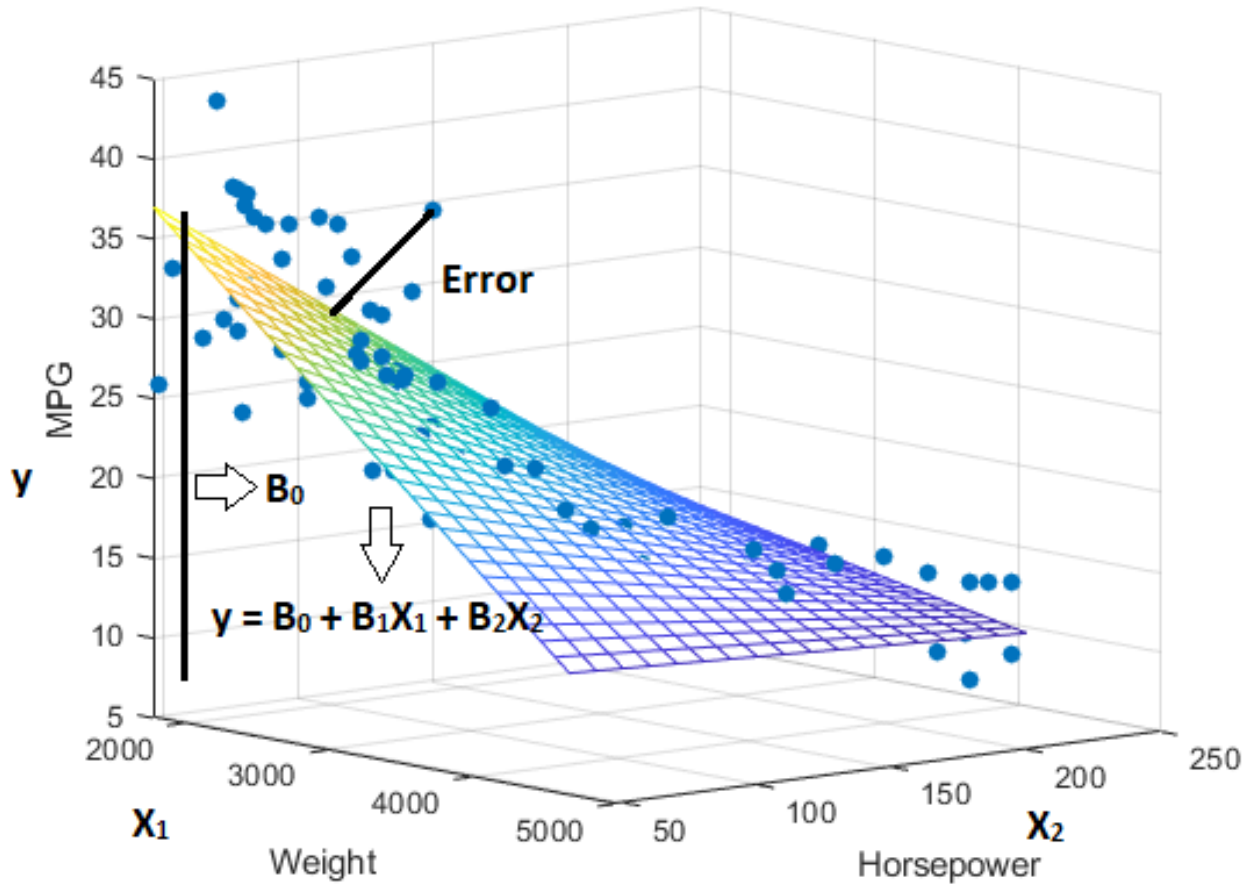
$y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \ldots\ldots + B_nX_n$

B_0 = y intercept

$B_1, B_2, \ldots B_n$ = slope for each independent variable

y = dependent variable

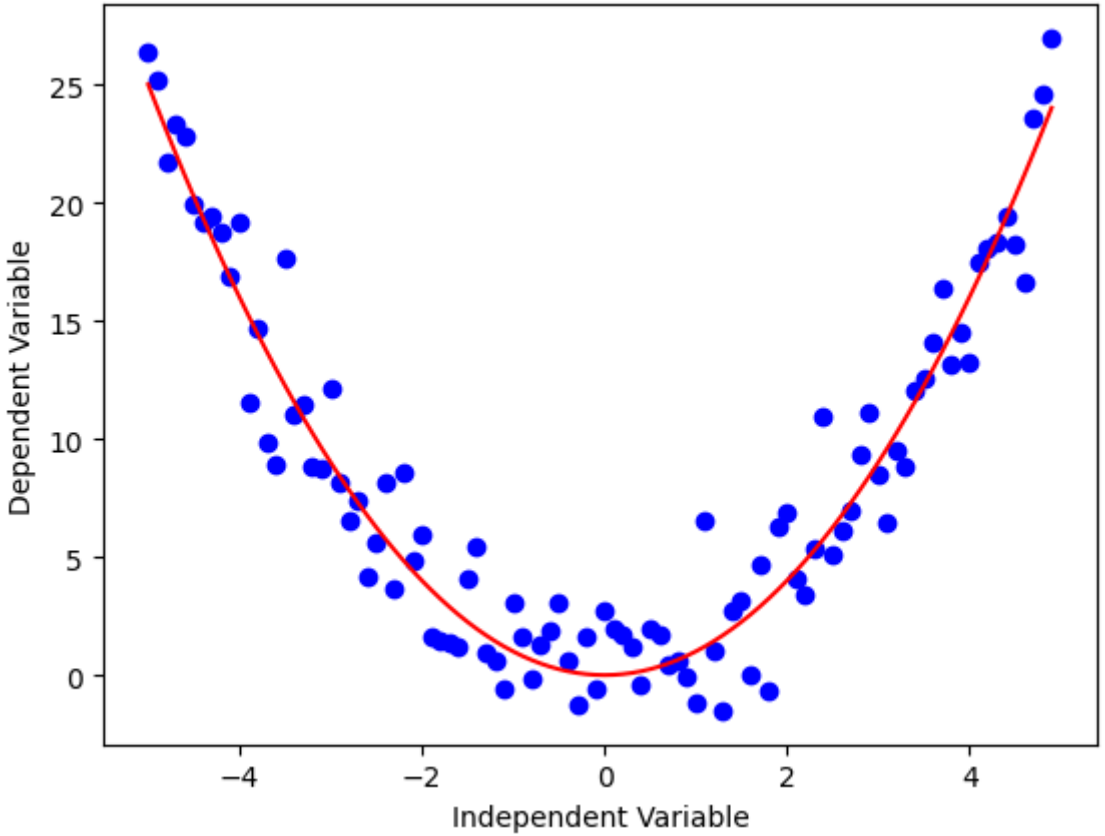
$X_1, X_2, \ldots, X_n$ = independent variables


45
40
35
30
25
20
15
10
5
MPG
y
Error
B0
y = B0 + B1X1 + B2X2
2000
3000
4000
5000
X1
Weight
50
100
150
200
250
X2
Horsepower

Nonlinear Regression:

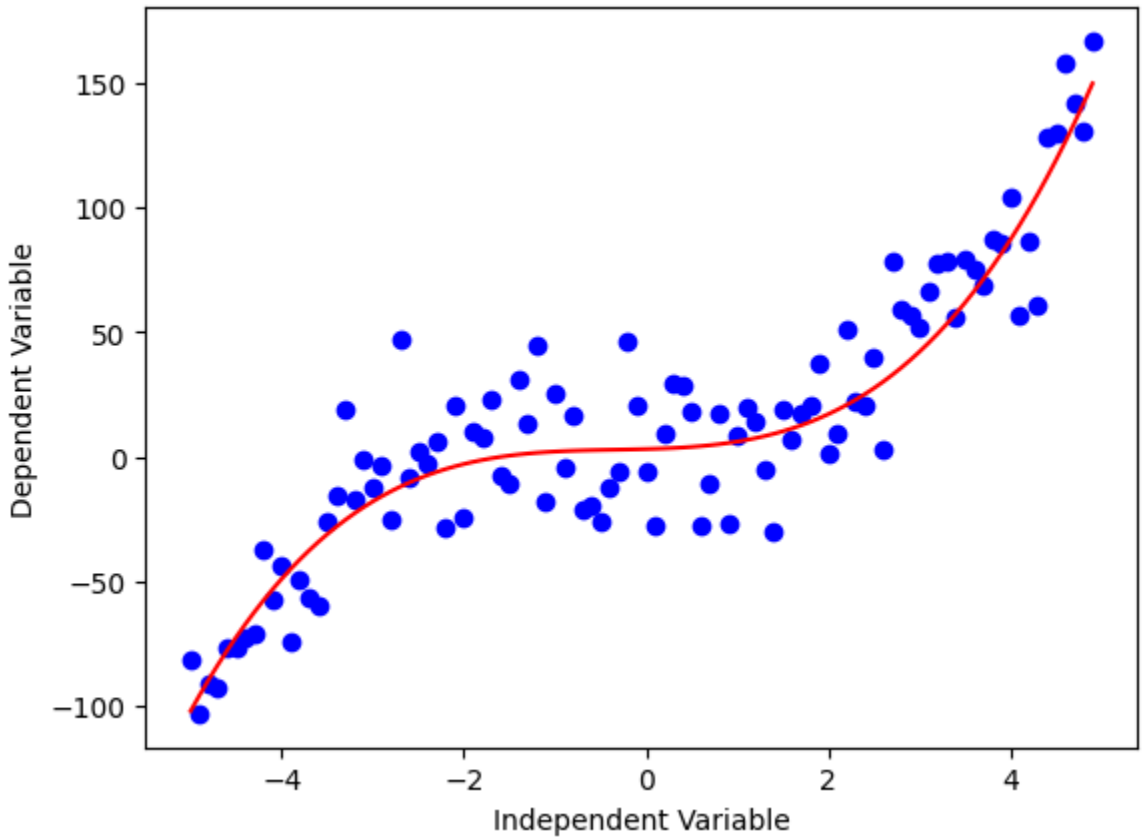
Non-Linear regression is a type of polynomial regression. It is a method to model a non-linear relationship between the dependent and independent variables. It is used in place when the data shows a curvy trend

$y = B_0 + B_1X + B_2X^2 + \ldots.. + B_NX^n$



Quadratic Nonlinear regression

Cubic Regression



Probability

Random experiment:

A random experiment is a type of experiment that has multiple possible outcomes. Such an experiment can be repeated many times.

Ex. Tossing a coin, Rolling a dice, Picking an object

Sample space (set):

Set of all possible outcome of a random experiment

S1={H,T} (Tossing a coin)

S2={1,2,3,4,5,6} (Rolling a dice)

Event:

Subset of sample space

E1={2,4,6} (Even number while rolling a dice)

E2={HH,HT,TH} (At least one head while tossing two coins)

The probability is the measure of the likelihood of an event to happen. It measures the certainty of the event. The formula for probability is given by;

- **P(E) = Number of Favorable Outcomes/Number of total outcomes**
- **P(E) = n(E)/n(S)**

Here,

- n(E) = Number of event favorable to event E
- n(S) = Total number of outcomes

Some important types of events:

Independent Events and Dependent Events

If the occurrence of any event is completely unaffected by the occurrence of any other event, such events are known as an **independent event** in probability and the events which are affected by other events are known as **dependent events**

Mutually Exclusive Events

If the occurrence of one event excludes the occurrence of another event, such events are mutually **exclusive events** i.e. two events don't have any common point. For example, if S = {1 , 2 , 3 , 4 , 5 , 6} and E_1, E_2 are two events such that E_1 consists of numbers less than 3 and E_2 consists of numbers greater than 4.

So, E1 = {1,2} and E2 = {5,6} .

Then, E1 and E2 are mutually exclusive.

Exhaustive Events

A set of events is called **exhaustive** if all the events together consume the entire sample space.

Complementary Events

For any event E_1 there exists another event E_1' which represents the remaining elements of the sample space S.

$E_1 = S - E_1$'

If a dice is rolled then the sample space S is given as S = {1 , 2 , 3 , 4 , 5 , 6 }. If event E_1 represents all the outcomes which is greater than 4, then E_1 = {5, 6} and E_1' = {1, 2, 3, 4}.

Thus E_1' is the complement of the event E_1.

Similarly, the complement of E_1, E_2, E_3…….E_n will be represented as E_1', E_2', E_3'…….E_n'

Events Associated with "OR"

If two events E_1 and E_2 are associated with **OR** then it means that either E_1 or E_2 or both. The union symbol **(U)** is used to represent OR in probability.

Thus, the event $E_1 U E_2$ denotes E_1 OR E_2.

If we have mutually exhaustive events $E_1, E_2, E_3 \ldots\ldots E_n$ associated with sample space S then,

$E_1 U E_2 U E_3 U \ldots\ldots E_n = S$

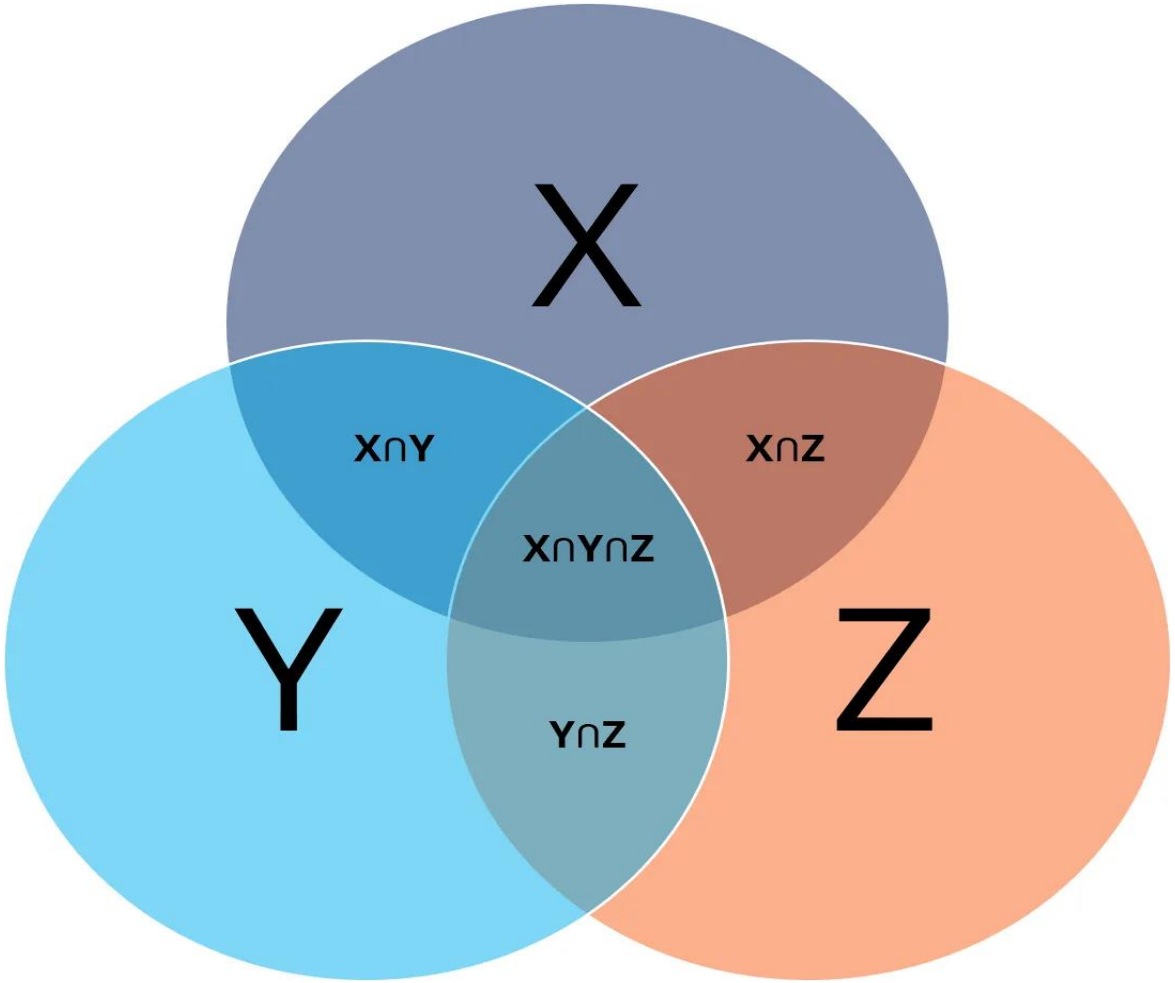
Events Associated with "AND"

If two events E_1 and E_2 are associated with **AND** then it means the intersection of elements which is common to both the events. The intersection symbol **(∩)** is used to represent AND in probability.

Thus, the event $E_1 \cap E_2$ denotes E_1 and E_2

Event E_1 but not E_2

It represents the difference between both the events. Event E_1 but not E_2 represents all the outcomes which are present in E_1 but not in E_2. Thus, the event E_1 but not E_2 is represented as $E_1 - E_2$



Conditional Probability

Let's take an example to understand this concept.

Example:

You are rolling a dice. What is the probability of getting prime number given that an even number showed up?

Solution:

Event E1 = getting even number.

Event E2 = getting prime number

We need to find out P(E2/E1) , that means probability of E2, consider E1 has already occurred.

P(E2/E1) = 1/3 (out of 2,4,6, only 2 is the prime number)

We can rewrite this as P(E2/E1) = $\frac{1/6}{3/6} = \frac{P(E1 \cap E2)}{P(E1)}$

P(E2/E1) = $\frac{P(E1 \cap E2)}{P(E1)}$ ---------- Conditional probability formula

$P(E1 \cap E2)$ = P(E2/E1) * P(E1) ------- equation 1

Let's assume the opposite scenario

What is the probability of getting even number given that a prime number showed up?

Event E1 = getting even number.

Event E2 = getting prime number

We need to find out P(E1/E2) , that means probability of E1, consider E2 has already occurred.

P(E1/E2) = 1/3 (out of 2,3,5, only 2 is the even number)

We can rewrite this as P(E1/E2) = $\frac{1/6}{3/6} = \frac{P(E1 \cap E2)}{P(E2)}$

P(E1/E2) = $\frac{P(E1 \cap E2)}{P(E2)}$

$P(E1 \cap E2)$ = P(E1/E2) * P(E2) ------- equation 2

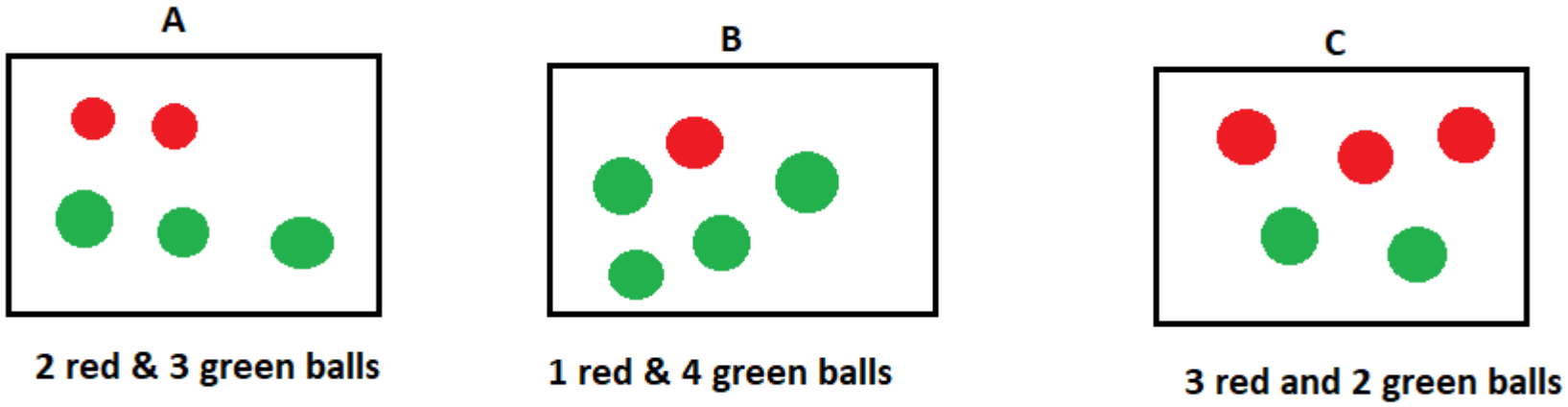
From equation 1 and equation 2, we get

$P(E2/E1) * P(E1) = P(E1/E2) * P(E2)$

$P(E2/E1) = \frac{P(E1/E2) * P(E2)}{P(E1)}$ -------this is an important equation to understand the Bayes theorem

Bayes Theorem:

Let's understand this with an example



A) What is the probability of getting a red ball given that A is chosen?

Answer: P(R/A) = 2/5

B) What is the probability of getting a red ball from bag A

Answer: $P(A \cap R) = P(A) * P(R/A)$

C) What is the probability of getting red ball?

Answer: $P(R) = P(A \cap R) + P(B \cap R) + P(C \cap R)$

D) What is the conditional probability that bag A is chosen given that red ball is drawn? (This is a Bayes theory problem)

Answer: $$P(A/R) = \frac{P(A \cap R)}{P(A \cap R) + P(B \cap R) + P(C \cap R)}$$

$$= \frac{P(A) * P(R/A)}{P(A) * P(R/A) + P(B) * P(R/B) + P(C) * P(R/C)}$$

$$P(A/R) = \frac{P(A) * P(R/A)}{P(R)}$$ (Bayes theorem)

$$P(A/R) = \frac{1/3 * 2/5}{(1/3 * 2/5)+(1/3 * 1/5)+(1/3 * 3/5)} = 1/3$$

So the conditional probability that bag A is chosen given that red ball is drawn, is 1/3.