STATISTICS

SAI SRAVYA TURAGA

February 2023

Table of Contents

CHAPTER 1 INTRODUCTION

1.1 Statistics and its types

It is the idea we can learn about the properties of large sets of objects or events (a population) by studying the characteristics of a smaller number of similar objects or events (a sample). Because in many cases gathering comprehensive data about an entire population is too costly, difficult, or flat out impossible, statistics start with a sample that can conveniently or affordably be observed.

Statistics makes work easy and simple and provides a clear and clean picture of work you do on a regular basis.

Two types of statistics

1. Descriptive statistics:
2. Inferential statistics:

In **Descriptive Statistics**, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean or standard deviation. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorised into four different categories:

1. Measure of frequency (given by the number of times a particular data occurs)
2. Measure of dispersion (Range, Variance, Standard Deviation)
3. Measure of central tendency (mean, median and mode)
4. Measure of position (percentile and quartile ranks.)

The **Inferential statistics** is used to interpret the meaning of Descriptive statistics. we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences to a population. It grants us permission to give

statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.

1.2 DATA ANALYTICS

It is the practice of examining data to answer questions, identify trends, and extract insights. 4 key types of data analytics

- Descriptive Analytics ("What happened?")
- Diagnostic Analytics ("Why did this happen?")
- Predictive Analytics ("What might happen in the future?")
- Prescriptive Analytics ("What should we do next?")

- **Descriptive Analytics**:

It allows you to pull trends from raw data and succinctly describe what happened or is currently happening.

For example, imagine you're analyzing your company's data and find there's a seasonal surge in sales for one of your products: a video game console. Here, descriptive analytics can tell you, "This video game console experiences an increase in sales in October, November, and early December each year."

Data visualization is a natural fit for communicating descriptive analysis because charts, graphs, and maps can show trends in data—as well as dips and spikes—in a clear, easily understandable way.

- **Diagnostic Analytics:**

This type includes comparing coexisting trends or movement, uncovering correlations between variables, and determining causal relationships where possible. Continuing the aforementioned example, you may dig into video game console users' demographic data and find that they're between the ages of eight and 18. The customers, however, tend to be between the ages of 35

and 55. Analysis of customer survey data reveals that one primary motivator for customers to purchase the video game console is to gift it to their children. The spike in sales in the fall and early winter months may be due to the holidays that include gift-giving.

Diagnostic analytics is useful for getting at the root of an organizational issue.

- **Predictive Analytics:**

It is used to make predictions about future trends or events. By analyzing historical data in tandem with industry trends, you can make informed predictions about what the future could hold for your company.

For instance, knowing that video game console sales have spiked in October, November, and early December every year for the past decade provides you with ample data to predict that the same trend will occur next year. Backed by upward trends in the video game industry, this is a reasonable prediction to make.

Making predictions for the future can help your organization formulate strategies based on likely scenarios.

- **Prescriptive Analytics:**

It takes into account all possible factors in a scenario and suggests actionable takeaways. This type of analytics can be especially useful when making data-driven decisions.

Rounding out the video game example: What should your team decide to do given the predicted trend in seasonality due to winter gift-giving? Perhaps you decide to run an A/B test with two ads: one that caters to product end-users (children) and one targeted to customers (their parents). The data from that test can inform how to capitalize on the seasonal spike and its supposed cause even further. Or, maybe you decide to increase marketing efforts in September with holiday-themed messaging to try to extend the spike into another month.

While manual prescriptive analysis is doable and accessible, machine-learning algorithms are often employed to help parse through large volumes of data to recommend the optimal next step.

1.3 Sampling methods

Two major types of Sampling, random and non-random

1.3.1 Random sampling:

In this kind of sampling, each element of the population has the same probability of getting selected in the sample.

- **Simple random sampling with replacement**: In simple random sampling with replacement, for the creation of a sample size n, you select an element from the population and then return it to the population. This procedure is repeated n times. Thus, each element of the population can be selected more than once in a sample. This is used when the population size is small.
- **Simple random sampling without replacement**: In simple random sampling without replacement, for the creation of a sample size n, you select an element from the population and don't return it to the population. The selection of elements from the population is repeated n times. This is used when the population size is large.
- **Stratified random sampling**: In stratified random sampling, the population is divided into strata on the basis of common characteristics. The elements are then selected from these strata.
- **Cluster sampling** : In cluster sampling, the population is divided into clusters, and then, a simple random sample of these clusters is selected.
- **Systematic sampling**: In systematic sampling, a starting point is selected in the population, and then, the elements are selected at regular, fixed intervals.

1.3.2 Non-random sampling:

In this kind of sampling, each element of the population does not have the same probability of getting selected in the sample.

- **Convenience sampling**: In convenience sampling, the researcher selects the elements from the population on the basis of the convenient accessibility of these elements.
- **Judgemental sampling**: In judgemental sampling, the researcher selects the elements on the basis of his judgement and bias.
- **Quota sampling** : The population is divided into groups or quotas, on the basis of which you select the sample. Quota sampling is, to a certain extent, similar to random sampling; the sampling procedure is more or less the same in both the cases, except the quota is fixed in quota sampling. That is, you don't consider the entire population, just a section of it to create a quota.
- **Snowball sampling**: In the case of snowball sampling, a small sample is first selected, say a sample of five people. Then, each of the five members can suggest five names, and those five can suggest five more each. This creates a snowball effect.

Examples.

exit poll ---- random sample
RBI survery for households--- cluster sampling
Drug test --- stratified sampling

Q. What is the difference between stratified random sampling and cluster sampling?

A. In stratified random sampling, the whole population is divided into strata based on common characteristics, and then, elements are selected from each stratum. In cluster sampling, on the other hand, the whole population is divided into clusters, and then, some of the clusters are chosen randomly to create a sample.

1.3.3 Other sampling techniques

There are four types of sampling methods/techniques.

- **Random Sampling:** In this method, people in the sample are selected randomly. This is like randomly pulling names out of a hat.

Example: Suppose you want to find out the average internet usage per person in India. You just put the names of all the Indians in a hat and pull out 100 names at random, and then calculate the average internet usage of these 100 Indians.

- **Stratified Sampling**: Here, people are divided into subgroups and then selected randomly from those subgroups. But this is done in such a way that the final sample has the same proportions of these subgroups as the population.

Example: Again, suppose you want to find out the average internet usage per person in India. Note that 70% of Indians live in rural areas, and 30% live in urban areas. So, you would put the names of all the rural Indians in hat A and the names of all the urban Indians in hat B. Then, you'd pull 70 names out of hat A and 30 names out of hat B. Now, again, you'd have a sample of 100 Indians, but this time, your sample would be more representative of the population as its rural and urban proportions would be the same as that of the population.

- **Volunteer Sampling**: Here, your sample is composed of people who want to volunteer for the survey.

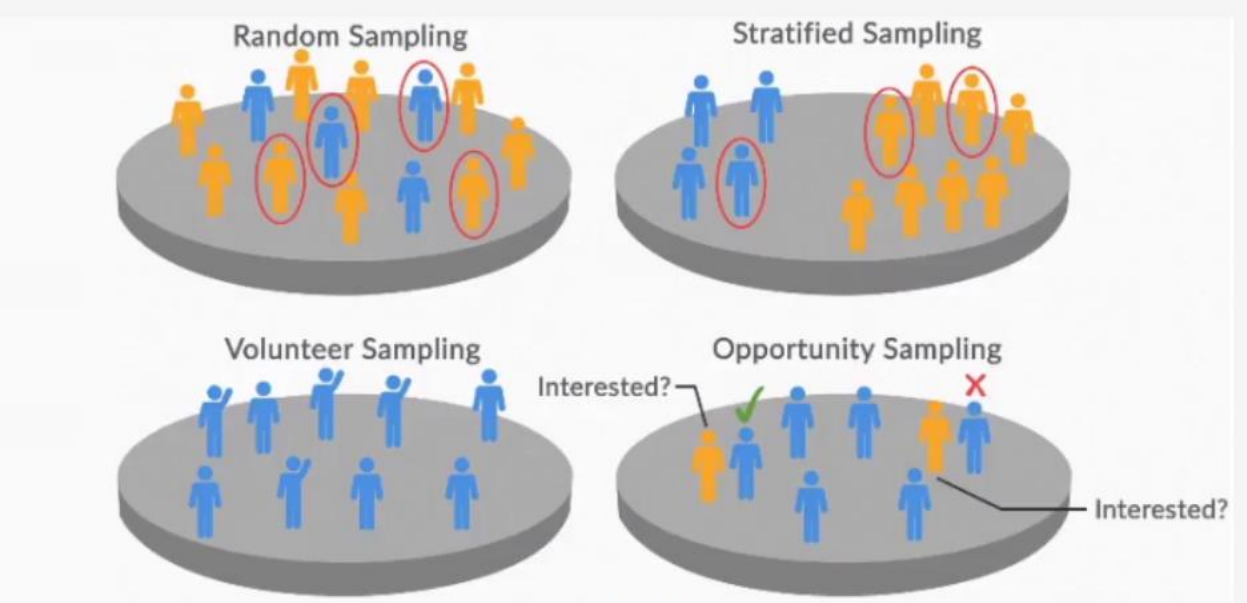
Example: Suppose that once more, you want to find out the average internet usage per person in India. You could ask people to take an online survey, which asks them how often/much they use the internet. You could ask the same question through a telephonic survey.

The good thing about this type of sampling is that it looks unbiased and random because the survey participants are selected at random through the medium (internet, telephone) itself. There is no human interference. However, the medium will also bring in some bias. For example, an internet survey is more likely to include people who have high internet usage, whereas a telephone survey is a little more likely to have a balanced representation of heavy internet users and people who use the internet infrequently.

- **Opportunity Sampling**: In this method, the people around and close to the surveyor form their sample space.

Example: This time, when you want to find out the average internet usage per person in India, you just ask 100 people around you about their internet usage.

Clearly, this sampling method has the potential to become extremely biased. The only good thing here, probably, is that this is a relatively convenient sampling method.



Types of Sampling Methods

https://www.andrews.edu/~calkins/math/edrm611/edrm01.htm#INFER

1.4 Important Static Terms:

- **Population (N):** It is actually a collection of set of individuals or objects or events whose properties are to be analyzed.
- **Sample(n):** A sample is a portion of a population selected for further analysis.
- **Parameter**: A parameter is a characteristic of the whole population
- **Satistic**: A statistic is a characteristic of a sample, presumably measurable.

Note: Parameter is to Population as Statistic is to Sample

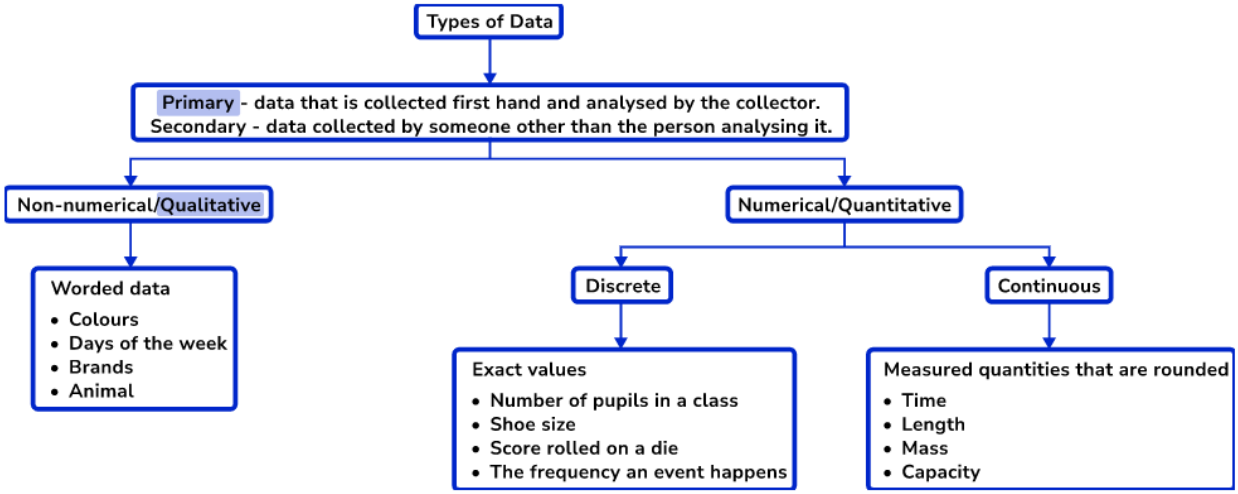
- **Data :** Its is facts or piecies of information that can be measured. Quantitative and qualitative

Data is of two types quantitative and Qualitative

- **Qualitative:** Qualitative data are nonnumeric. Qualitative data are often termed catagorical data

eg : {Poor, Fair, Good, Better, Best}, colors (ignoring any physical causes), and types of material {straw, sticks, bricks}

- **Quantitative:** Quantitative data are numeric. They are further classified as either discrete or continuous.
    - **Discrete data** are numeric data that have a finite number of possible values.eg. finite subset of the counting numbers, {1,2,3,4,5}
    - **Continuous Data:** Continuous data have infinite possibilities: 1.4, 1.41, 1.414, 1.4142, 1.141421... The real numbers are continuous with no gaps or interruptions. Physically measureable quantities of length, volume, time, mass, etc. are generally considered continuous.



1.5 Levels of Measurement:

Level of measurement or scale of measure is a classification that describes the nature of information within the values assigned to variables

- **Nominal**: Nominal data have no order and thus only gives names or labels to various categories. eg: gender, marital status, college major, and blood type.

- **Ordinal**: Ordinal data have order, but the interval between measurements is not meaningful.

eg: Education level (primary, secondary, post-secondary). Income (low, middle, and high). Overall status (poor to excellent). Likert and other scales of agreement (strongly disagree to strongly agree). Rank (such as sporting teams and class standings).

- **Interval**: Interval data have meaningful intervals between measurements, but there is no true starting point (zero).

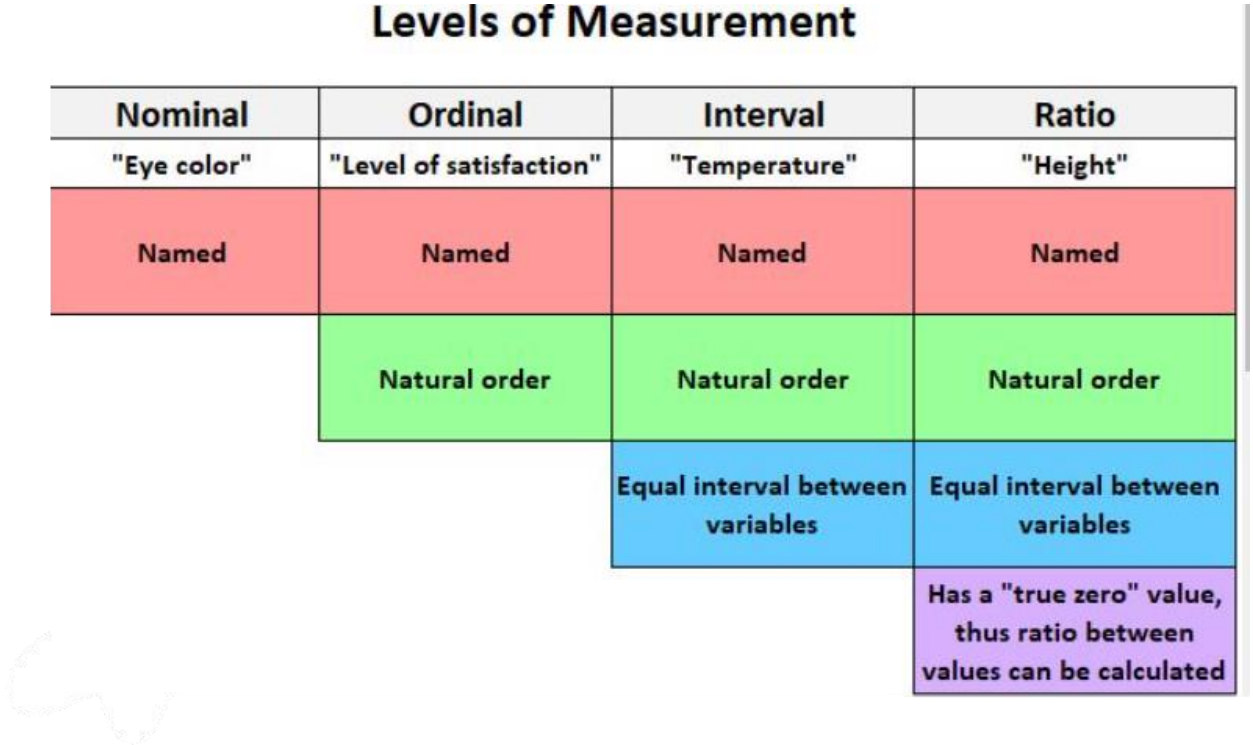
eg : Test scores (e.g., IQ or exams) ,(0 in exam doesn't mean 0 iq)

Temperature in Fahrenheit or Celsius (0 C doen't mean 0 temperature)

- **Ratio**: Ratio data have the highest level of measurement. Ratios between measurements as well as intervals are meaningful because there is a starting point (zero). eg : Height, Age,Weight,Temperature in Kelvin

Nominal: the data can only be categorised
Ordinal: the data can be categorised and ranked
Interval: the data can be categorised, ranked, and evenly spaced
Ratio: the data can be categorised, ranked, evenly spaced, and has a natural zero.

https://www.scribbr.co.uk/stats/measurement-levels/



1.6 Measure of central Tendency:

These measures are used to represent the typical value or center point of any data set.

- **Mean:**

The arithmetic mean is the average of a group of numbers and is computed by summing all numbers and dividing by the number of numbers.

The mean summarizes an entire dataset with a single number representing the data's center point.The mean gives us an idea of where the "center" of a dataset is located.

Population Mean(μ): $$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + x_3 + \ldots + x_N}{N}$$

Sample Mean($\bar{x}$): $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

- **Median:**

The middle value of the ordered array of numbers. For an array with an odd number of terms, the median is the middle number. For an array with an even number of terms, the median is the average of the two middle numbers. eg: Salaries

- **Mode:**

The most frequent value in the dataset. If the data have multiple values that occurred the most frequently, we have a multimodal distribution (data sets with more than 2 modes).

1.7 Measure of Position

It is a method by which the position that a particular data value has within a given data set can be identified. Measures of position are techniques that divide a set of data into equal groups. To determine the measurement of position, the data must be sort from lowest to highest.

- **Percentiles**: A percentile is a number where a certain percentage of scores fall below that number. For example, a 90th percentile marks the spot where 90% of values fall below that cut-off point.

- **Quartiles**: Simply put, quartiles divide your data into quarters: the lowest quarter, two middle quarters, and a highest quarter.

- **Deciles**: Deciles are like quartiles. But where quartiles split the data in four equal parts, deciles split the data into ten parts: The 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, 90th and 100th percentiles.

- **Five Number Summary**: It is an overview of your data. It includes 5 items:

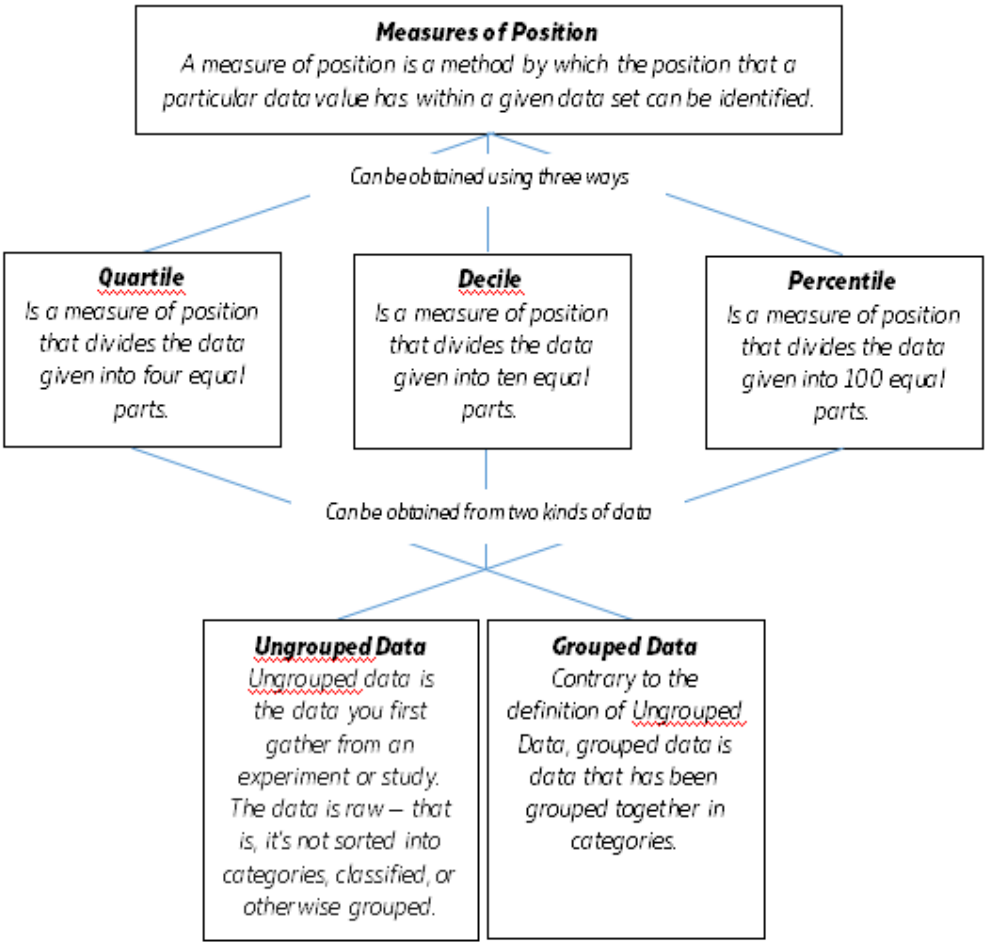
    - The minimum.
    - Q1 (the first quartile, or the 25% mark).
    - The median.
    - Q3 (the third quartile, or the 75% mark).
    - The maximum.

- **Interquartile Range (IQR):** The interquartile range tells you where the "middle fifty" is in a data set. While the range tells you where the beginning and end are in a set, the IQR shows you where the bulk of the "middling" values lie.

- **Outliers :** Outliers are unusual values that fall outside of an expected range of values. For example, if you're measuring IQ values of children, your statistics would be thrown off if Einstein and Stephen Hawking were in your class: their IQs would be outliers.

- **Box and Whiskers Plot**: It shows the spread and center of data. It is a graphical representation of the five number summary.

- **Standard scores (i.e. z-scores)**: Z-scores are a way to compare results from a test to a "normal" population.



1.8 Measure of Dispersion

In statistics, the measures of dispersion help to interpret the variability(spread) of data i.e. to know how much homogenous or heterogeneous the data is. In simple terms, it shows how squeezed or scattered the variable is.

Types of Measures of Dispersion: There are two main types of dispersion methods in statistics which are:

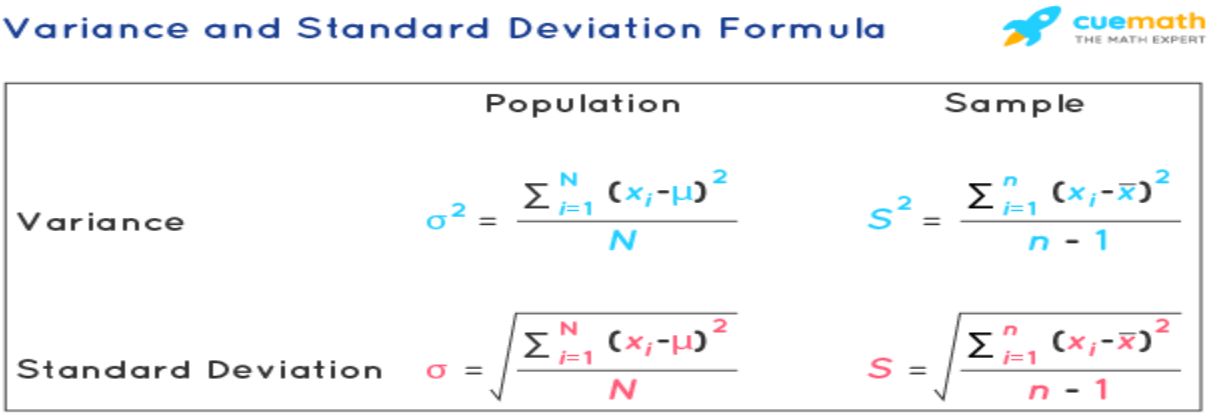
- Absolute Measure of Dispersion
- Relative Measure of Dispersion

Absolute Measure of Dispersion:

An absolute measure of dispersion contains the same unit as the original data set. The absolute dispersion method expresses the variations in terms of the average of deviations of observations like standard or means deviations. It includes range, standard deviation, quartile deviation, etc.

The types of absolute measures of dispersion are:

1. **Range:** It is simply the difference between the maximum value and the minimum value given in a data set. Example: 1, 3,5, 6, 7 => Range = 7 -1= 6
2. **Variance:** It is the expected value of the squared variation of a random variable from its mean value, in probability and statistics. Informally, variance estimates how far a set of numbers (random) are spread out from their mean value.



Note: Sample variance is divided by (n-1) to correct the bias because we are using the *sample* mean, instead of the *population* mean, to calculate the variance. Since the sample mean is based on the data, it will get drawn toward the center of mass for the data..

- **Standard Deviation:** The square root of the variance is known as the standard deviation i.e. S.D. = √σ.

Both measures reflect variability in a distribution, but their units differ: SD is expressed in the same units as the original values (e.g., minutes or meters).Variance is expressed in much larger units (e.g., meters squared)

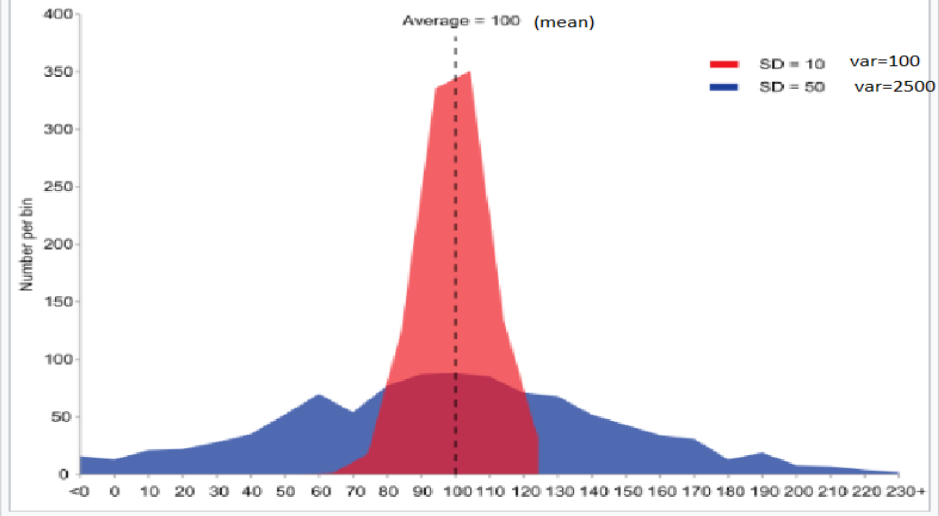


Figure 1 Sample with same mean but different variance

3. **Quartiles and Quartile Deviation(Q3-Q1):** The quartiles are values that divide a list of numbers into quarters. The quartile deviation is half of the distance between the third and the first quartile.
4. **Mean and Mean Deviation:** The average of numbers is known as the mean and the arithmetic mean of the absolute deviations of the observations from a measure of central tendency is known as the mean deviation (also called mean absolute deviation).

Relative Measure of Dispersion

The relative measures of dispersion are used to compare the distribution of two or more data sets. This measure compares values without units. Common relative dispersion methods include:

- Co-efficient of Range
- Co-efficient of Variation
- Co-efficient of Standard Deviation
- Co-efficient of Quartile Deviation
- Co-efficient of Mean Deviation

Co-efficient of Dispersion

The coefficients of dispersion are calculated (along with the measure of dispersion) when two series are compared, that differ widely in their averages. The dispersion coefficient is also used when two series with different measurement units are compared. It is denoted as C.D.

The common coefficients of dispersion are:

C.D. in terms of	Coefficient of dispersion
Range	C.D. = (Xmax – Xmin) / (Xmax + Xmin)
Quartile Deviation	C.D. = (Q3 – Q1) / (Q3 + Q1)
Standard Deviation (S.D.)	C.D. = S.D. / Mean
Mean Deviation	C.D. = Mean deviation/Average

- **Standard Deviation Vs Variance:**

Both standard deviation and variance used to see how data is spread but standard deviation gives more clarity about the deviation of data from a mean.
Variance mostly used Anova & There are different use case where either one of them suits
Variance Example: you might want to understand how much variance in test scores can be explained by IQ and how much variance can be explained by hours studied.If 36% of the variation is due to IQ and 64% is due to hours studied, that's easy to understand.
But if we use the standard deviations of 6 and 8, that's much less intuitive and doesn't make much sense in the context of the problem.
Standard Deviation Example: If we want to see how the volume of perfume bottles is manufactured then in this case if stdev is 2 then we can say 95% bottles volume will be between +- 2 stddev. but variance doesn't make sense here.

Important Note: Stddev is in the same unit as original data, but variance is squared so stddev helps more when you compare the result with original data. Both stddev as well as variance are affected by outliers

- **Z-Score (standard score):**

The value of the z-score tells you how many standard deviations the datapoint is away from the mean. In other words, the Z-score tells us how far away a particular observation is from the average value of a dataset, measured in terms of the number of standard deviations.

A positive z-score indicates the raw score is higher than the mean average. For example, if a z-score is equal to +1, it is 1 standard deviation above the mean. A negative z-score reveals the raw score is below the mean average.

$$z = (x-\mu)/\sigma$$

In order to use a z-score, we need to know the population mean (μ) and also the population standard deviation (σ). Depending upon whether the given Z-Score is positive or negative, one makes use of the respective positive Z-Table or negative Z-Table .

The Z-score is often used in statistical analysis to identify outliers or unusual data points in a dataset. It is also useful for comparing data points from different datasets that have different scales or units of measurement, as the Z-score standardizes the data by putting them on a common scale.

To identify outliers using Z-score, we can follow these steps:

1. Calculate the mean (μ) and standard deviation (σ) of the dataset.
2. Calculate the Z-score of each observation using the formula: Z = (x - μ) / σ, where x is the observation value.
3. Identify any observation with a Z-score greater than a certain threshold. A common threshold is a Z-score of 2.5 or 3, which corresponds to being more than 2.5 or 3

standard deviations away from the mean.

Question:

You take the GATE examination and score 500. The mean score for the GATE is 390 and the standard deviation is 45. How well did you score on the test compared to the average test taker?

Solution:

The following data is readily available in the above question statement

Raw score/observed value = X = 500

Mean score = μ = 390

Standard deviation = σ = 45

By applying the formula of z-score,

$$z = (X - \mu) / \sigma$$
$$z = (500 - 390) / 45$$
$$z = 110 / 45 = 2.44$$

*This means that your z-score is **2.44**.*

Since the Z-Score is positive 2.44, we will make use of the positive Z-Table.

Now let's take a look at Z Table (CC-BY) to know how well you scored compared to the other test-takers.

Follow the instruction below to find the probability from the table.

*Here, **z-score = 2.44***

1. *Firstly, map the first two digits 2.4 on the Y-axis.*
2. *Then along the X-axis, map 0.04*
3. *Join both axes. The intersection of the two will provide you the Z-Score value you're looking for*

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558
2	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972

*As a result, you will get the final value which is **0.99266**.*

Now, we need to compare how our original score of 500 on the GATE examination compares to the average score of the batch. To do that we need to convert the Z-Score into a percentage value.

0.99266 * 100 = 99.266%

*Finally, you can say that you have performed well than almost **99%** of other test-takers.*

Question:

What is the probability that a student scores between 350 and 400 (with a mean score μ of 390 and a standard deviation σ of 45)?

Solution:

Min score = X1 = 350

Max score = X2 = 400

By applying the formula of z-score,

$$z1 = (X1 - \mu) / \sigma$$
$$z1 = (350 - 390) / 45$$
$$z1 = -40 / 45 = -0.88$$

$$z2 = (X2 - \mu) / \sigma$$
$$z2 = (400 - 390) / 45$$
$$z2 = 10 / 45 = 0.22$$

Since z1 is negative, we will have to look at a negative Z-Table and find that p1, the first probability, is ***0.18943****.*

z2 is positive, so we use a positive Z-Table which yields a probability p2 of ***0.58706****.*

The final probability is computed by subtracting p1 from p2: :

$$p = p2 - p1$$
$$p = 0.58706 - 0.18943 = 0.39763$$

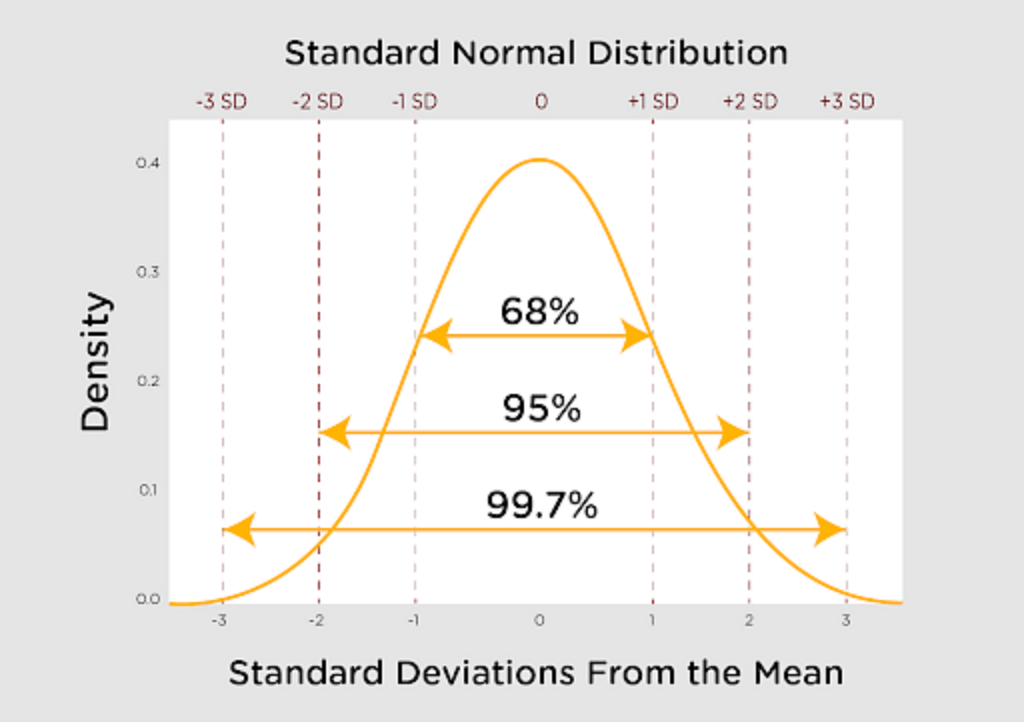
The probability that a student scores between 350 and 400 is ***39.763%*** *(0.39763 * 100).*

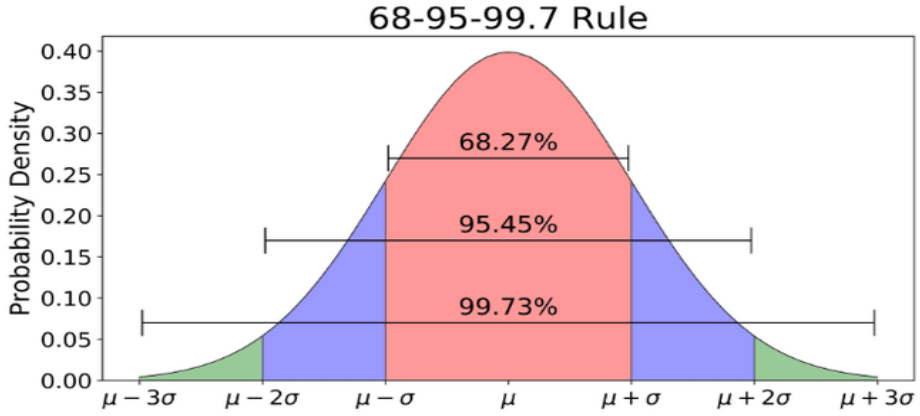
Two ways of applying the standard deviation are the empirical rule and Chebyshev's theorem based on data being normally distributed(mean=mode=median) or not

- **Empirical Rule**

If a set of data is normally distributed, or bell shaped, approximately 68% of the data values are within one standard deviation of the mean, 95% are within two standard deviations, and almost 100% are within three standard deviations.

In terms of zscore, If the number of elements in a given set is large, then about 68% of the elements have a z-score between -1 and 1; about 95% have a z-score between -2 and 2; about 99% have a z-score between -3 and 3. This is known as the Empirical Rule or the 68-95-99.7 Rule and can be demonstrated in the image below.





The 68–95–99.7 Rule for a Normal Distribution

Suppose a recent report states that for California, the average statewide price of a gallon of regular gasoline is \$3.12. Suppose regular gasoline prices vary across the state with a standard deviation of \$0.08 and are normally distributed. According to the empirical rule, approximately 68% of the prices should fall within $\mu \pm 1\sigma$, or \$3.12 ± 1 (\$0.08). Approximately 68% of the prices should be between \$3.04 and \$3.20, as shown in Figure 3.5A. Approximately 95% should fall within $\mu \pm 2\sigma$ or \$3.12 ± 2 (\$0.08) = \$3.12 ± \$0.16, or between \$2.96 and \$3.28, as shown in Figure 3.5B. Nearly all regular gasoline prices (99.7%) should fall between \$2.88 and \$3.36 ($\mu \pm 3\sigma$).

- **Chebyshev's Theorem:**

Chebyshev's theorem applies to all distributions regardless of their shape and thus can be used whenever the data distribution shape is unknown or is non-normal. Even though Chebyshev's theorem can in theory be applied to data that are normally distributed.
Chebyshev's theorem states that at least 1-(1/k^2) values will fall within k + or - K standarddeviations of the mean regardless of the shape of the distribution.**(2stdev -75% & 3stdev - 88%)**

CHEBYSHEV'S THEOREM

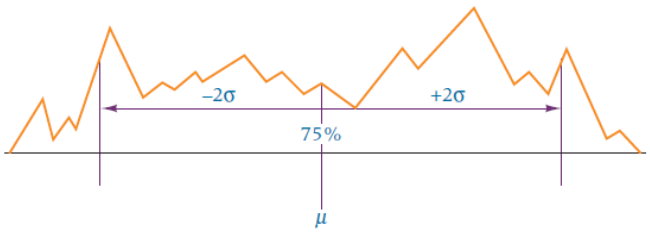
Within k standard deviations of the mean, $\mu \pm k\sigma$, lie at least

$$1 - \frac{1}{k^2}$$

proportion of the values.
Assumption: $k > 1$

Specifically, Chebyshev's theorem says that at least 75% of all values are within $\pm 2\sigma$ of the mean regardless of the shape of a distribution because if $k = 2$, then $1 - 1/k^2 = 1 - 1/2^2 = 3/4 = .75$. Figure 3.6 provides a graphic illustration. In contrast, the empirical

FIGURE 3.6
Application of Chebyshev's Theorem for Two Standard Deviations



- **Covariance:**

Covariance is a measure of how much two random variables varytogether. covariance doesn't tell us the strength of a relationship and it only tells us how the relationship is whether it is +ve or -ve or not. Covariance is a statistical tool that is used to determine the relationship between the movement of two asset prices. When two stocks tend to move together, they are seen as having a positive **covariance**; when they move inversely, the **covariance** is negative.

- **Correlation:**

Measure the relationship between two variables and range from *-1 to 1*, the normalized version of covariance.it tells us the strength of the relationship.

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$cor(x,y) = \frac{cov(x,y)}{\sqrt{var(x)\ var(y)}}$$

1.9 Skewness and Kurtosis

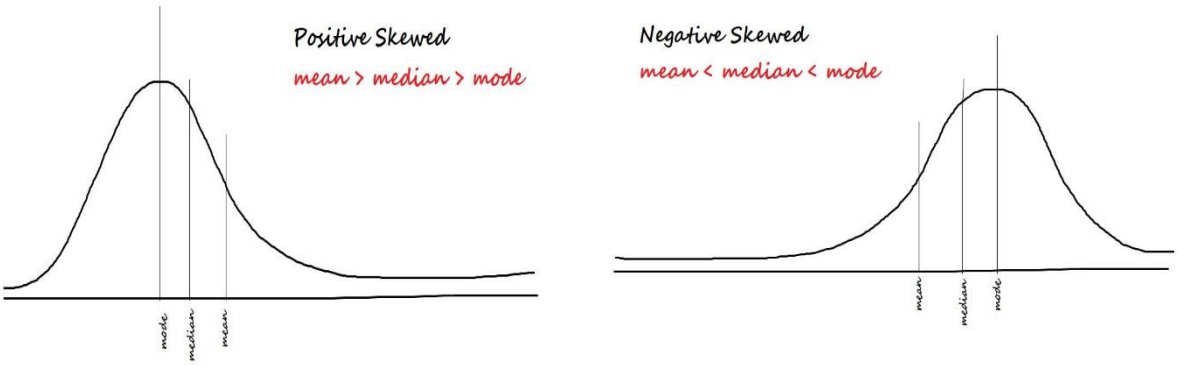
Skewness and Kurtosis are used to describe the spread and height of your normal distribution. Skewness is used to denote the horizontal pull on the data.

Skewness :

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

A standard normal distribution is perfectly symmetrical and has zero skew. Other examples of zero-skewed distributions are the T Distribution, the uniform distribution and the Laplace distribution. However, other distributions don't have zero skew. The two types of Skewness are:

- **Positive/right-skewed:** Data is said to be positively skewed if most of the data is concentrated to the left side and has a tail towards the right.

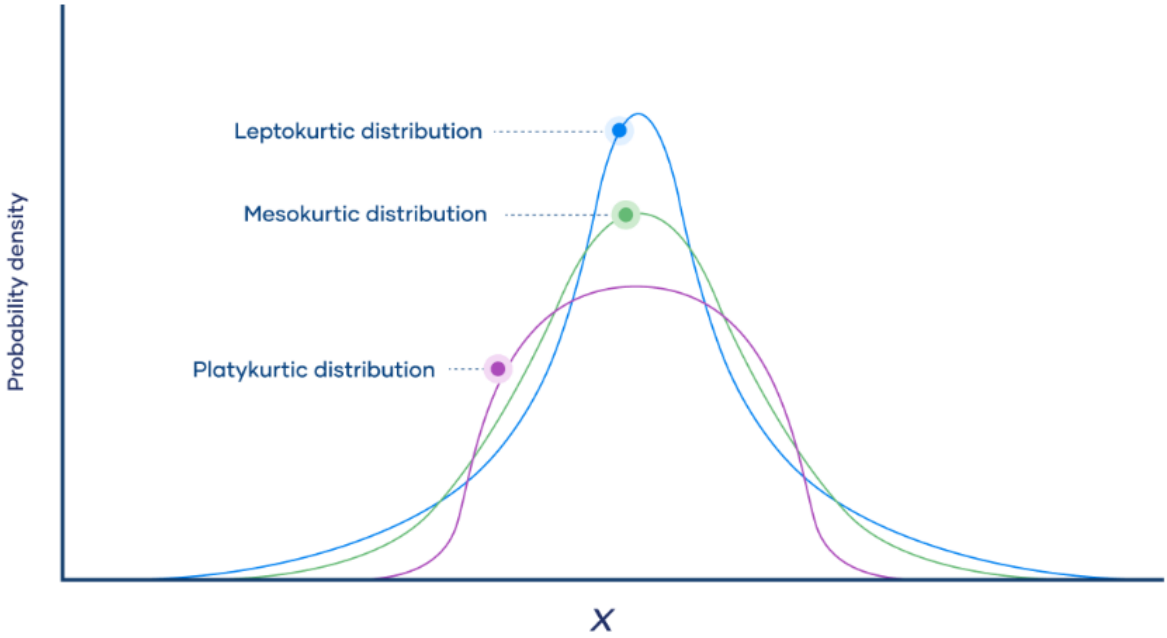


- **Negative/left-skewed:** Data is said to be negatively skewed if most of the data is concentrated to the right side and has a tail towards the left.

Skewness is a measure of symmetry in a distribution. Actually, it's more correct to describe it as a measure of **lack** of symmetry. A standard normal distribution is perfectly symmetrical and has zero skew. Other examples of zero-skewed distributions are the T Distribution, the uniform distribution and the Laplace distribution. However, other distributions don't have zero skew. Therefore, we need a way to calculate how much the distribution is skewed.

- **Kurtosis:**

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case. The histogram is an effective graphical technique for showing both the skewness and kurtosis of data set.



- Distributions with medium kurtosis (medium tails) are **mesokurtic**.
- Distributions with low kurtosis (thin tails) are **platykurtic**.
- Distributions with high kurtosis (fat tails) are **leptokurtic**.

Tails are the tapering ends on either side of a distribution. They represent the probability or frequency of values that are extremely high or low compared to the mean. In other words, tails represent how often **outliers** occur.

	Mesokurtic	Platykurtic	Leptokurtic
Tailedness	Medium-tailed	Thin-tailed	Fat-tailed
Outlier frequency	Medium	Low	High
Kurtosis	Moderate(3)	Low(<3)	High(>3)
Excess Kurtosis	0	Negative	Positive
Example distribution	Normal	Uniform	Laplace

Since normal distributions have a kurtosis of 3, excess kurtosis is calculated by subtracting kurtosis by 3. S = standard deviation, n= number of data points, mean= $\bar{X}$

Excess kurtosis = Kurt – 3

$$Sample\ Skewness = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}$$

$$Sample\ Kurtosis = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{s^4}$$

CHAPTER 2 Probability

The word 'Probability' means the chance of occurrence of a particular event.

Probability of event to happen P(E)= $\frac{Number\ of\ favourable\ outcomes}{Total\ Number\ of\ outcomes}$

2.1 *Basics of probability:* ***Link***

Term	Definition	Example
Sample Space	The set of all the possible outcomes to occur in any trial	1. Tossing a coin, Sample Space (S) = {H,T} 2. Rolling a die, Sample Space (S) = {1,2,3,4,5,6}
Sample Point	It is one of the possible results	In a deck of Cards: • 4 of hearts is a sample point. • The queen of clubs is a sample point.
Experiment or Trial	A series of actions where the outcomes are always uncertain.	The tossing of a coin, Selecting a card from a deck of cards, throwing a dice.
Event	It is a single outcome of an experiment.	Getting a Heads while tossing a coin is an event.
Outcome	Possible result of a trial/experiment	T (tail) is a possible outcome when a coin is tossed.
Complimentary event	The non-happening events. The complement of an event A is the event, not A (or A')	In a standard 52-card deck, A = Draw a heart, then A' = Don't draw a heart
Odds	Odds is usually defined in statistics as the **probability an event will occur divided by the probability that it will not occur**. In other words, it's a ratio of successes (or wins) to losses (or failures). As an example.	, if a racehorse runs 100 races and wins 20 times, the odds of the horse winning a race is 20/80 = 1/4
Impossible Event	The event cannot happen	In tossing a coin, impossible to get both head and tail at the same time

Term	Definition	Example
Equally Likely Events	Events are said to be equally likely if one of them cannot be expected to occur in preference to others. In other words, it means each outcome is as likely to occur as any other outcome.	When a die is thrown, all the six faces, i.e., 1, 2, 3, 4, 5 and 6 are equally likely to occur.
Mutually exclusive events or Disjoint Events	Events are called mutually exclusive if they cannot occur simultaneously.	In a deck of 52 cards, drawing a red card and drawing a club are mutually exclusive events because all the clubs are black.
Independent Events:	Events A and B are said to be independent if the occurrence of any one event does not affect the occurrence of any other event.	
Dependent Event:	Events are said to be dependent if occurrence of one affects the occurrence of other events.	
exhaustive events	The set of events out of which one will definitely occur whenever the experiment is performed is called exhaustive events in probability. The union of the exhaustive events gives the entire sample space.	Sample space S = {1, 2, 3, 4, 5, 6} • A be the event of getting a number greater than 3 . A = {4, 5, 6} • B be the event of getting a number greater than 2 but less than 5. B = {3, 4} • C be the event of getting a number less than 3. C = {1, 2} then A ∪ B ∪ C = {4, 5, 6} ∪ {3, 4} ∪ {1, 2} = {1, 2, 3, 4, 5, 6} = S Therefore, A, B, and C are called exhaustive events. However, the probability of exhaustive events is equal to 1.

2.2 Multiplication Rule:

To find the probability of two events occurring in sequence, you can use the Multiplication Rule. Before applying multiplication first check whether event is dependent or independent then
based on this use following formulas:
1] The probability that two dependent events A and B will occur in sequence is

P(A and B) = P(A) * P(B | A) .

2] The probability that two independent events A and B will occur in sequence is

P**(A and B) = P(A) * P(B) .**

Ex-Two cards are selected, without replacing the first card, from a standard deck of 52 playing cards. Find the probability of selecting a king and then selecting a queen?
Sol. Because the first card is not replaced, the events are dependent.
P(K and Q) = P(K) * P(Q | K)
= 4 / 52 * 4 / 51= 16 / 2652≈ 0.006

Ex. A coin is tossed and a die is rolled. Find the probability of tossing a head and then rolling a 6.
Sol. The events are independent so,
P(H and 6) = P(H) * P(6)= 1 / 2 * 1 / 6= 1 / 12≈ 0.083

2.3 Bayes Theorem:

Bayes' Theorem states that the conditional probability of an event, based on the occu rrence of another event, is equal.

$$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)} = \frac{P(A).P(\frac{B}{A})}{P(B)}$$

Which tells us: how often A happens *given that B happens*, written **P(A|B)**,
When we know: how often B happens *given that A happens*, written **P(B|A)**
and how likely A is on its own, written **P(A)**
and how likely B is on its own, written **P(B)**

Example: Picnic Day

You are planning a picnic today, but the morning is cloudy

- Oh no! 50% of all rainy days start off cloudy!
- But cloudy mornings are common (about 40% of days start cloudy)
- And this is usually a dry month (only 3 of 30 days tend to be rainy, or 10%)

What is the chance of rain during the day?

We will use Rain to mean rain during the day, and Cloud to mean cloudy morning.

The chance of Rain given Cloud is written P(Rain|Cloud)

So let's put that in the formula:

$$P(\text{Rain}|\text{Cloud}) = \frac{P(\text{Rain})\ P(\text{Cloud}|\text{Rain})}{P(\text{Cloud})}$$

- P(Rain) is Probability of Rain = 10%
- P(Cloud|Rain) is Probability of Cloud, given that Rain happens = 50%
- P(Cloud) is Probability of Cloud = 40%

$$P(\text{Rain}|\text{Cloud}) = \frac{0.1 \times 0.5}{0.4} = .125$$

Or a 12.5% chance of rain. Not too bad, let's have a picnic!

2.4 Addition Rule:

The probability that events A or B will occur.

P(A or B) = P(A) + P(B) - P(A and B)

If events A and B are mutually exclusive, then the rule can be simplified to:

P(A or B) = P(A) + P(B)

EX- A card that is a 4 cannot be an ace. So, the events are mutually exclusive, as shown in the Venn diagram. The probability of selecting a 4 or an ace is:
P(4) or ace =P(4)+P(ace)=4/52+4/52=2/13

2.5 Permutations:

A permutation is an ordered arrangement of objects. The number of different permutations of n distinct objects is n!.

$$_nP_r = \frac{n!}{(n-r)!},$$

Note: In permutation order is important

EX-Find the number of ways of forming four-digit codes in which no digit is repeated

$$\begin{aligned} {}_nP_r &= {}_{10}P_4 \\ &= \frac{10!}{(10-4)!} \\ &= \frac{10!}{6!} \\ &= \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6!}{6!} \\ &= 5040 \end{aligned}$$

2.6 Combinations:

The number of combinations of r objects selected from a group of n objects without regard to order.

$$_nC_r = \frac{n!}{(n-r)!r!}$$

Note:In Combination order does not important

EX-Find the probability of being dealt 5 diamonds from a standard deck of 52 playing cards.
In a standard deck of playing cards, 13 cards are diamonds. Note that it does not matter what order the cards are selected. The possible number of ways of choosing 5 diamonds out of 13 is 13C5. The number of possible five-card hands is 52C5. So, the probability of being dealt 5 diamonds is

$$\begin{aligned} P(5 \text{ diamonds}) &= \frac{{}_{13}C_5}{{}_{52}C_5} \\ &= \frac{1287}{2{,}598{,}960} \\ &\approx 0.0005. \end{aligned}$$

2.7 Law of large numbers:

The law of large numbers is a principle of probability according to which the frequencies of events with the same likelihood of occurrence even out, given enough trials or instances. As the number of experiments increases, the actual ratio of outcomes will converge on the theoretical, or expected, ratio of outcomes.

For example, if a fair coin (where heads and tails come up equally often) is tossed 1,000,000 times, about half of the tosses will come up heads, and half will come up tails. The heads-to-tails ratio will be extremely close to 1:1. However, if the same coin is tossed only 10 times, the ratio will likely not be 1:1, and in fact might come out far different, say 3:7 or even 0:10.

CHAPTER 3 Inferential Statistics and Probability Distributions

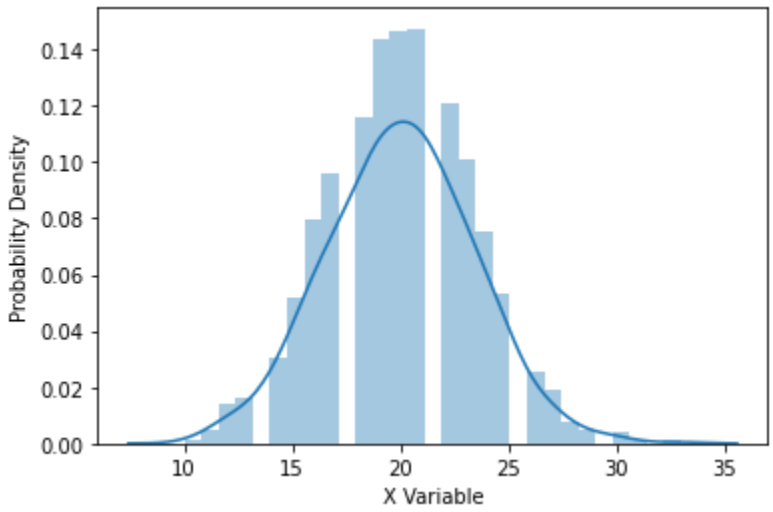
Source

3.1 Concepts of Statistics and Probability

3.1.1 Types of Distributions

Python code for important distributions: Link

Let X be a random variable that has more than one possible outcome. Plot the probability on the y-axis and the outcome on the x-axis. If we repeat the experiment many times and plot the probability of each possible outcome, we get a plot that represents the probabilities. This plot is called the probability distribution (PD). The height of the graph for X gives the probability of that outcome.



Why should you know about statistical distributions?
Because it helps in:

- Cleaning data and identifying anomalies/outliers
- Mapping to known characteristics of physical, social, economical, and natural phenomena

Distributions are of 2 types :

- **Continuous probability distributions:** These distributions model the probabilities of random variables that can have any possible outcome. For example, the possible values for the random variable X that represents weights of citizens in a town which can have any value like 34.5, 47.7, etc.,
  Common Continuous Distributions:
  - Normal/Gaussian Distribution
  - Student's t-Distribution
  - Uniform Distribution
  - Chi-Square
  - Exponential

- For a continuous variable the probability of getting exact value of x is always 0

  $P(X<=175.4)=P(X=175.4)+ P(X<175.4)= 0+ P(X<175.4)= P(X<175.4)$

- **Discrete probability distributions:** These distributions model the probabilities of random variables that can have discrete values as outcomes. For example, the possible values for the random variable X that represents the number of heads that can occur when a coin is tossed twice are the set {0, 1, 2} and not any value from 0 to 2 like 0.1 or 1.6.

  Common Discrete Distributions:

  - Binomial Distribution
  - Bernoulli Distribution
  - Uniform Distribution
  - Poisson Distribution
  - Hypergeometric
  - Negative Binomial

Each PD provides us extra information on the behavior of the data involved. Each PD is given by a probability function that generalizes the probabilities of the outcomes.

3.1.2 Probability Mass function (PMF) and Probability Density function (PDF)

Using this, we can estimate the probability of a particular outcome(discrete) or the chance that it lies within a particular range of values for any given outcome(continuous). The function is called a **Probability Mass function (PMF)** for discrete distributions and a **Probability Density function** (PDF) for continuous distributions. The total value of PMF and PDF over the entire domain is always equal to one.

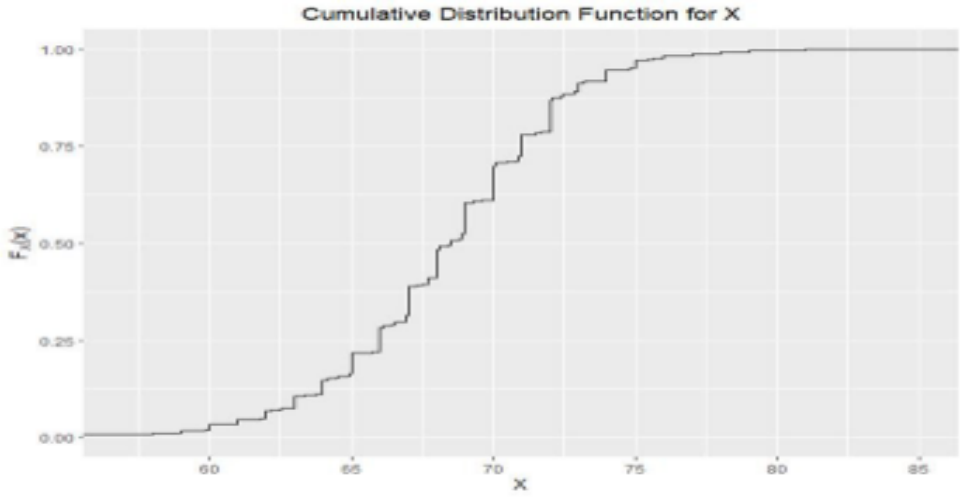
3.1.3 Cumulative Distribution Function

The PDF gives the probability of a particular outcome whereas the Cumulative Distribution Function gives the probability of seeing an outcome less than or equal to a particular value of the random variable. CDFs are used to check how the probability has added up to a certain point. For example, if P(X = 5) is the probability that the number of heads on flipping a coin is 5 then, P(X <= 5) denotes the cumulative probability of obtaining 1 to 5 heads.

Cumulative distribution functions are also used to calculate p-values as a part of performing hypothesis testing.

The Continuous Case

The CDF of continuous random variables is noisier than that of discrete variables. Just as previously done, we sum up (rather say, integrate) the PDF to get the CDF. For the example of the height, the following CDF has been made:



Note : PDF's are much used in real life compared to CDF's as the patterns, trends are much visible in PDF's .

Note : Area under PDF curve gives us cumulative probability.

3.1.4 Expected Value & Standard Deviation of Discrete random Variable:

Mean or Expectation or Expected Value: Source

The ***mean*** of a discrete random variable *X* is a weighted average of the possible values that the random variable can take. Unlike the sample mean of a group of observations, which gives each observation equal weight, the mean of a random variable weights each outcome x_i according to its probability, p_i. The common symbol for the mean (also known as the ***expected value*** of *X*) is μ, formally defined by

$$\mu = E(x) = \sum x p(x)$$

$$\mu_x = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$

$$= \sum x_i p_i$$

Variance and Standard Deviation:

The **variance** of a discrete random variable is

$$\sigma^2 = \Sigma(x - \mu)^2 P(x).$$

The **standard deviation** is

$$\sigma = \sqrt{\sigma^2} = \sqrt{\Sigma(x - \mu)^2 P(x)}.$$

Important Note: All the casino games are designed in such a way that they ensure a negative(-ve) expected value.

Casino example for expected value:

Suppose an individual plays a gambling game where it is possible to lose $1.00, break even, win $3.00, or win $5.00 each time she plays. The probability distribution for each outcome is provided by the following table:

Outcome	-$1.00	$0.00	$3.00	$5.00
Probability	0.30	0.40	0.20	0.10

The mean outcome for this game is calculated as follows:

μ = (-1*.3) + (0*.4) + (3*.2) + (10*0.1) = -0.3 + 0.6 + 0.5 = 0.8.

In the long run, then, the player can expect to win about 80 cents playing this game -- the odds are in her favor.

- In the above gambling game, suppose the casino realizes that it is losing money in the long term and decides to adjust the payout levels by subtracting $1.00 from each prize. The new probability distribution for each outcome is provided by the following table:

Outcome	-$2.00	-$1.00	$2.00	$4.00
Probability	0.30	0.40	0.20	0.10

The new mean is (-2*0.3) + (-1*0.4) + (2*0.2) + (4*0.1) = -0.6 + -0.4 + 0.4 + 0.4 = -0.2. This is equivalent to subtracting $1.00 from the original value of the mean, 0.8 -1.00 = -0.2. With the new payouts, the casino can expect to win 20 cents in the long run.

- Suppose that the casino decides that the game does not have an impressive enough top prize with the lower payouts, and decides to double all of the prizes, as follows:

Outcome	-$4.00	-$2.00	$4.00	$8.00
Probability	0.30	0.40	0.20	0.10

Now the mean is (-4*0.3) + (-2*0.4) + (4*0.2) + (8*0.1) = -1.2 + -0.8 + 0.8 + 0.8 = -0.4. This is equivalent to multiplying the previous value of the mean by 2, increasing the expected winnings of the casino to 40 cents.

Overall, the difference between the original value of the mean (0.8) and the new value of the mean (-0.4) may be summarized by (0.8 - 1.0)*2 = -0.4.

3.1.5 Properties of Means

- If a random variable *X* is adjusted by multiplying by the value *b* and adding the value *a*, then the mean is affected as follows:

$$\mu_{a+bX} = a + b\mu_X$$

- The mean of the sum of two random variables *X* and *Y* is the sum of their means**:** $\mu_{X+Y} = \mu_X + \mu_Y$

For example, suppose a casino offers one gambling game whose mean winnings are -$0.20 per play, and another game whose mean winnings are -$0.10 per play. Then the mean winnings for an individual simultaneously playing both games per play are -$0.20 + -$0.10 = -$0.30.

3.1.6 Central Limit Theorem

The **sampling distribution**, which is basically the distribution of sample means of a population, has some interesting properties which are collectively called the **central limit theorem**, which states that no matter how the original population is distributed, the sampling distribution will follow these three properties –

1. **Sampling Distribution's Mean** (μx) = **Population Mean** (μ)
2. Sampling Distribution's Standard Deviation (**Standard Error**) = $\sigma / \sqrt{n}$, where σ is the population's standard deviation and n is the sample size
3. For **n > 30**, the sampling distribution becomes a **normal** distribution

3.1.7 Notations (Population/Sample):

Population/Sample	Term	Notation	Formula
Population $(X_1, X_2, X_3, \ldots\ldots, X_N)$	Population Size	N	Number of items/elements in the population
	Population Mean	μ	$\frac{\sum_{i=1}^{i=N} X_i}{N}$
	Population Variance	σ^2	$\frac{\sum_{i=1}^{i=N}(X_i - \mu)^2}{N}$
Sample $(X_1, X_2, X_3, \ldots\ldots, X_n)$ (Sample of Population)	Sample Size	n	Number of items/elements in the sample
	Sample Mean	$\bar{X}$	$\frac{\sum_{i=1}^{i=n} X_i}{n}$
	Sample Variance	S^2	$\frac{\sum_{i=1}^{i=n}(X_i - \bar{X})^2}{n-1}$
Sampling Distribution of the Sample Mean $(\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots\ldots, \bar{X}_k)$ (k Sample Means)	Sampling Distribution's Size	No convention (We have used k, but that is not a norm)	
	Sampling Distribution's Mean (mean of sample means)	$\mu_{\bar{X}}$	$\mu_{\bar{X}} = \mu$
	Sampling Distribution's Standard Deviation	S.E. (Standard Error)	S.E. = $\sigma/\sqrt{n}$

3.1.8 Confidence Intervals (Mean estimation)

Using CLT, you can estimate the population mean from the sample mean and standard deviation. For example, to estimate the mean commute time of 30,000 employees of an office, you took a sample of 100 employees and found their mean commute time. For this sample, the sample mean $\bar{X}$ = 36.6 minutes, sample standard deviation S = 10 minutes.
Using CLT, you can say that the sampling distribution for mean commute time will have -
1. Mean = μ {unknown}
2. Standard error = $\sigma/\sqrt{n} = S/\sqrt{n} = 10/\sqrt{100}$ = 1
3. Since n(100) > 30, the sampling distribution is a normal distribution

Using these properties, you can **claim** that the probability that the population mean μ lies between 34.6 (36.6-2) mins and 38.6 (36.6+2) mins, is 95.4%.

Also, there is some terminology related to the claim -
1. Probability associated with the claim is called **confidence level** (Here it is 95.4%)
2. Maximum error made in sample mean is called **margin of error** (Here it is 2 minutes)
3. Final interval of values is called **confidence interval** {Here it is the range – (34.6, 38.6)}

In fact, you can generalise the entire process. Let's say you have a sample with sample size n, mean $\bar{X}$ and standard deviation S. Now, the y% confidence interval (i.e., confidence interval corresponding to y% confidence level) for μ will be given by the range—

$$\boldsymbol{Confidence\ Interval} = (\bar{\boldsymbol{X}} - \frac{\boldsymbol{Z^*S}}{\sqrt{\boldsymbol{n}}}, \bar{\boldsymbol{X}} + \frac{\boldsymbol{Z^*S}}{\sqrt{\boldsymbol{n}}})$$

Where, **Z*** is the Z-score associated with a y% confidence level.

For example, the 90% confidence interval for the mean commute time will be

$$\boldsymbol{\mu} = (\bar{\boldsymbol{X}} - \frac{\boldsymbol{Z^*S}}{\sqrt{\boldsymbol{n}}}, \bar{\boldsymbol{X}} + \frac{\boldsymbol{Z^*S}}{\sqrt{\boldsymbol{n}}})$$

Here,
$\bar{X} = 36.6$ minutes
S = 10 minutes
n = 100
Z* = 1.65 (Z* corresponding to 90% confidence level)
So, the confidence interval is – μ = (34.95 mins, 38.25 mins)

3.2 Discrete Probability Distributions

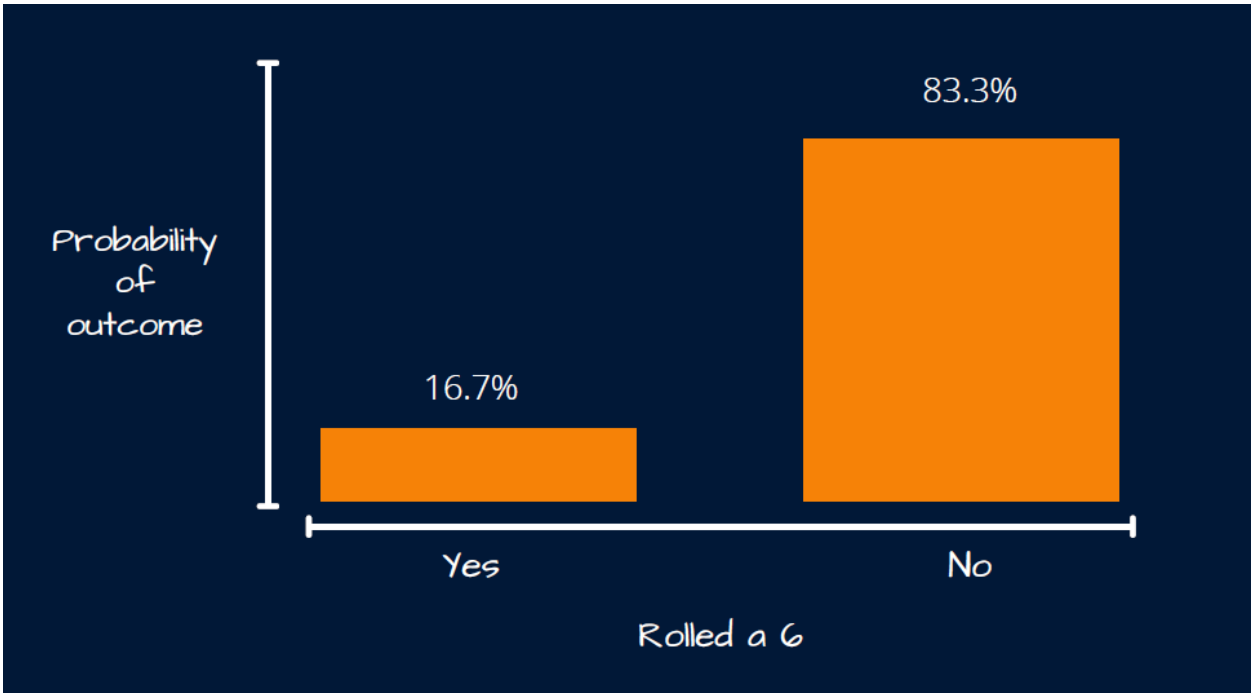
3.2.1 Bernoulli Distribution

This distribution is generated when we perform an experiment only once(1 trail) and it has only two possible outcomes – success and failure. A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment and a sequence of outcomes is called a Bernoulli process

$$f(x) = \begin{cases} p^x * (1-p)^{1-x} \\ 0 \end{cases} \quad \begin{matrix} if\ x = 0,1 \\ otherwise \end{matrix} = \begin{cases} p & if\ x = 1 \\ 1-p & if\ x = 0 \end{cases}$$

Example : Toss a Coin leads one time to determine win or lose. There is no intermediate result. The occurrence of a head denotes success, and the occurrence of a tail denotes failure

- Bernoulli distributions are used to view the probability that an investment will succeed or fail.
- If we roll a die many, many times, we should end up with a probability of rolling a 6, 1 out of every 6 times (or 16.7%) and thus a probability of not rolling a 6, in other words rolling a 1,2,3,4 or 5, 5 times out of 6 (or 83.3%) of the time!



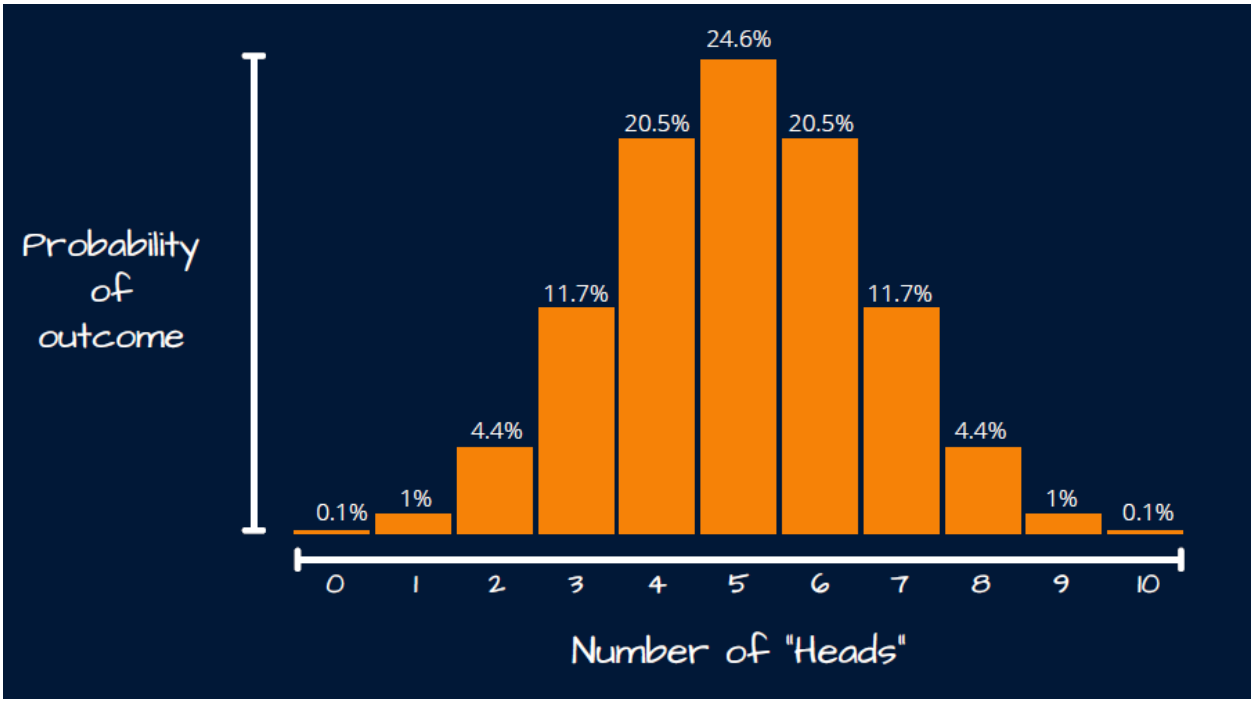
3.2.2 Binomial Distribution

In this distribution we have *n* independent and identical Bernoulli trials. This means that each trial does not influence other trials, they each have the same probability of success, and each trials has exactly two outcomes (success or failure). We then measure the number of successes over these *n* trials.

Ex: The binomial distribution is used in options pricing models that rely on binomial trees. In a binomial tree model, the underlying asset can only be worth exactly one of two possible

values—with the model, there are just two probable outcomes with each iteration—a move up or a move down with defined values.

A Binomial Distribution can end up looking a lot like the shape of a normal distribution. The main difference is that instead of plotting continuous data, it instead plots a distribution of **two possible discrete outcomes** for example, the results from flipping a coin.Imagine flipping a coin 10 times, and from those 10 flips, noting down how many were "Heads". It could be any number between 1 and 10.
Now imagine repeating that task 1,000 times...



If the coin we are using is indeed fair (not biased to heads or tails) then the distribution of outcomes should start to look the plot above. In the vast majority of cases we get 4, 5, or 6 "heads" from each set of 10 flips, and the likelihood of getting more extreme results is much more rare!

A binomial experiment is a probability experiment that satisfies these conditions:

- The experiment has a fixed number of trials, where each trial is independent of the othertrials.
- There are only two possible outcomes of interest for each trial. Each outcome can be classified as a success (S) or as a failure (F).
- The probability of success is the same for each trial.
- The random variable x counts the number of successful trials.

Notations:

n - The number of trials
p - The probability of success in a single trial
q - The probability of failure in a single trial q = (1 - p)
x - The random variable represents a count of the number of successes in n trials: x = 0, 1, 2, 3,..

Binomial Probability formula along with mean ,variance and std:

In a binomial experiment, the probability of exactly x successes in n trials is

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

Note that the number of failures is $n - x$.

$$\text{Mean: } \mu = np$$

$$\text{Variance: } \sigma^2 = npq$$

$$\text{Standard deviation: } \sigma = \sqrt{npq}$$

Example :

If 40% of all graduate business students at a large university are women and if random samples of 10 graduate business students are selected many times, the expectation is that, on average, four of the 10 students would be women.
Picking marbles from a bag containing red and green marbles with replacement

3.2.3 Multinomial Distributions:

Multinomial distributions occur when there is a probability of more than two outcomes with multiple counts. For instance, say you have a covered bowl with one green, one red, and one yellow marble. For your test, you record the number of times you randomly choose each of the marbles for your sample.
Ex : In finance and investing, these distributions estimate the probability that a specific set of financial events will occur.

3.2.4 Poisson Distribution

The Poisson distribution expresses the probability that a given number of events will occur over a fixed period.

The Poisson distribution describes the occurrence of **rare** events per some time interval. A Poisson experiment does not have a given number of trials (n) as a binomial experiment does. For example, a Poisson experiment might focus on the number of cars randomly arriving at an automobile repair facility during a 10-minute interval.
The Poisson distribution is a discrete probability distribution of a random variable x that satisfies these conditions.

- The experiment consists of counting the number of times x an event occurs in a given interval. The interval can be an interval of time, area, or volume.
- The probability of the event occurring is the same for each interval.

- The number of occurrences in one interval is independent of the number of occurrences in other intervals.

The probability of exactly x occurrences in an interval is

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

where e is an irrational number approximately equal to 2.71828 and μ is the mean number of occurrences per interval unit.
EX: Number of telephone calls per minute at a small business

The mean or expected value of a Poisson distribution is meu.

Example:
The mean number of accidents per month at a certain intersection is three. What is the probability that in any given month four accidents will occur at this intersection?
Using x = 4 and m = 3, the probability that 4 accidents will occur in any given month at the intersection is

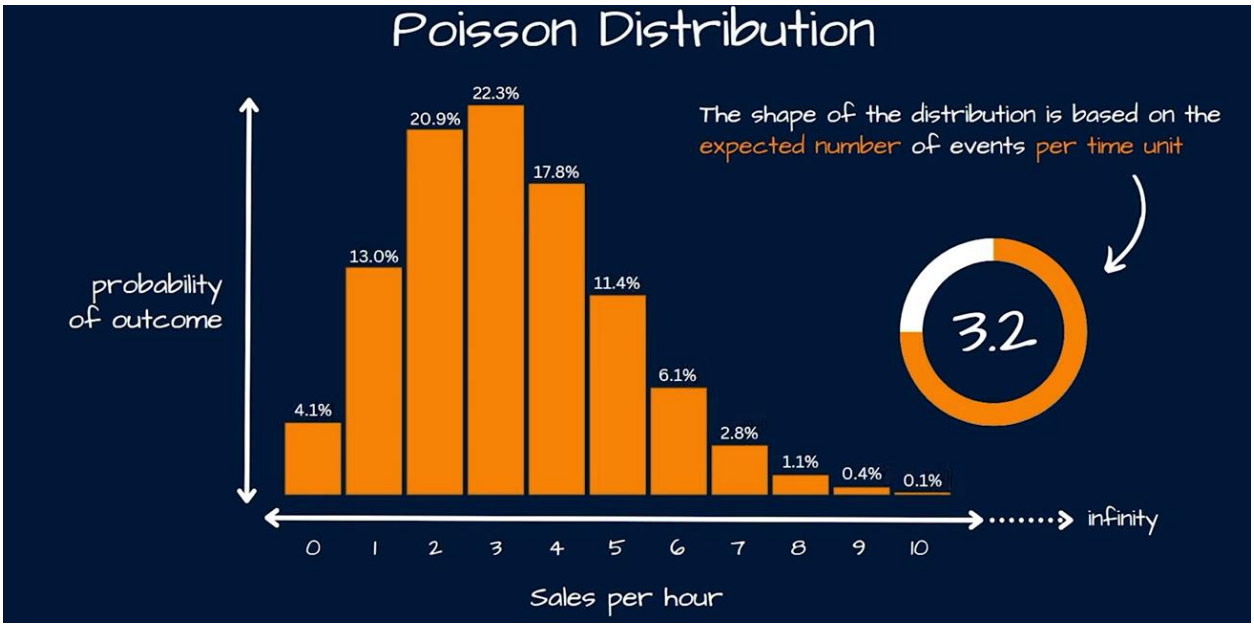
$$P(4) \approx \frac{3^4(2.71828)^{-3}}{4!} \approx 0.168.$$

The Poisson distribution is a discrete distribution that counts the frequency of occurrences as integers, whose list {0, 1, 2, ...} can be infinite. For instance, say you have a covered bowl with one red and one green marble, and your chosen period is two minutes. Your test is to record whether you pick the green or red marble, with the green indicating success. After each test, you place the marble back in the bowl and record the results.

In this model, the distribution would be plotting the results over a period of time, indicating how often green is chosen.

Poisson distribution is commonly used to model financial data where the tally is small and often zero. For example, it can be used to model the number of trades a typical investor will make in a given day, which can be 0 (often), 1, 2, and so on.

A **Poisson Distribution** is a **discrete** distribution similar to the Binomial Distribution (in that we're plotting the probability of whole numbered outcomes) Unlike the other distributions we have seen however, this one is **not symmetrical** - it is instead bounded between 0 and infinity The Poisson distribution describes the number of events or outcomes that occur during some fixed interval. Most commonly this is a time interval like in our example below where we are plotting the distribution of **sales per hour** in a shop.



3.2.5 Geometric Distribution:

Many actions in life are repeated until a success occurs. For instance, you might have to send an email several times before it is successfully sent. A situation such as this can be represented by a geometric distribution.

A geometric distribution is a discrete probability distribution of a random variable x that satisfies these conditions.

- A trial is repeated until a success occurs.
- The repeated trials are independent of each other.
- The probability of success p is the same for each trial.
- The random variable x represents the number of trials in which the first success occurs.

The probability that the first success will occur on trial number x is:

$$P(x) = pq^{x-1}, \text{ where } q = 1 - p.$$

In other words, when the first success occurs on the third trial, the outcome is FFS, and the probability is

$$P(3) = q \cdot q \cdot p, \text{ or } P(3) = p \cdot q^2.$$

3.2.6 Hypergeometric distribution

. The hypergeometric distribution applies only to experiments in which the trials are done without replacement.

The hypergeometric distribution, like the binomial distribution, consists of two possible outcomes: success and failure. However, the user must know the size of the population and the proportion of successes and failures in the population to apply the hypergeometric distribution.

In other words, because the hypergeometric distribution is used when sampling is done without replacement, information about population makeup must be known in order to redetermine the probability of a success in each successive trial as the probability changes.

3.3 Continuous Distributions

3.3.1 Normal Distribution (Gaussian Distribution)

A normal distribution is a continuous probability distribution for a random variable x.
The graph of a normal distribution is called the normal curve.

Properties.

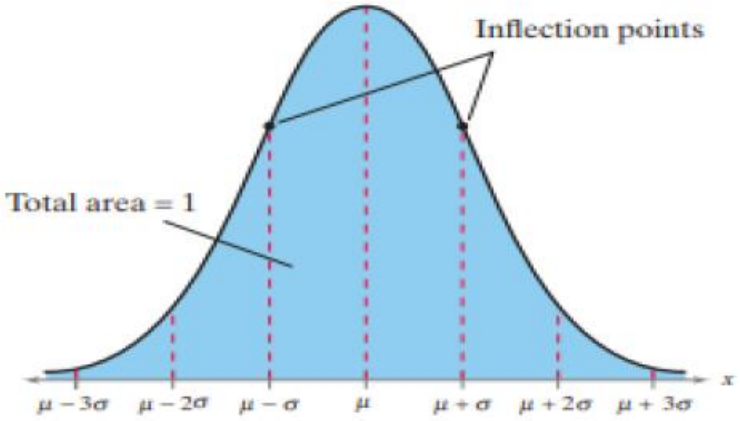
1. The mean, median, and mode are equal.
2. The normal curve is bell-shaped and is symmetric about the mean.
3. The total area under the normal curve is equal to 1.
4. The normal curve approaches, but never touches, the x-axis as it extends farther and farther away from the mean.

Standard Normal Distribution:
There are infinitely many normal distributions, each with its own mean and standard deviation. The normal distribution with a mean of 0 and a standard deviation of 1 is called the **standard normal distribution**. The horizontal scale of the graph of the standard normal distribution corresponds to z-scores. z-score is a measure of position that indicates the number of standard deviations a value lies from the mean.

$$z\_score = \frac{value - mean}{standard\ deviation} = \frac{x - \mu}{\sigma}$$

When each data value of a normally distributed random variable x is transformed into a z-score, the result will be the standard normal distribution.

1. The cumulative area is close to 0 for z-scores close to $z = -3.49$.
2. The cumulative area increases as the z-scores increase.
3. The cumulative area for $z = 0$ is 0.5000.
4. The cumulative area is close to 1 for z-scores close to $z = 3.49$.



When a random variable x is normally distributed, you can find the probability that x will lie in an interval by calculating the area under the normal curve for the interval. To find the area under any normal curve, first convert the upper and lower bounds of the interval to z-scores. Then use the standard normal distribution to find the area.

For example: heights, blood pressure, measurement error, and IQ scores follow the normal distribution.

Normal distribution vs Standard distribution.

A normal distribution can take on any value as its mean and standard deviation. On the otherhand, a standard normal distribution has always the fixed mean and standard deviation.

3.3.2 Uniform distribution

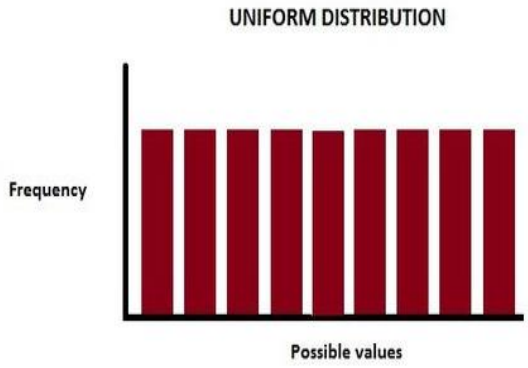
Uniform distribution gives the same probability to any points of a set. It has both discrete and continuous forms.

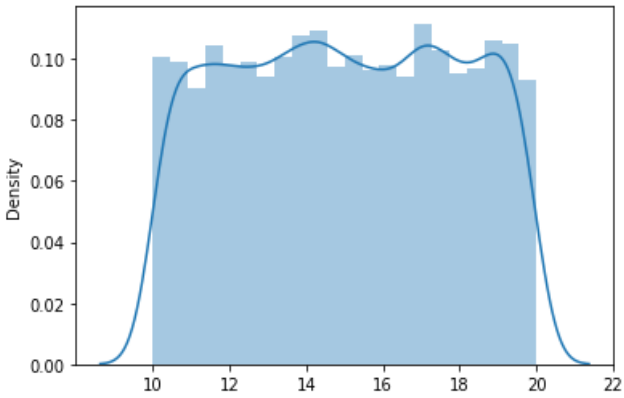
The uniform distribution is determined from a probability density function that contains equal values along some interval between the points a and b. Basically, the height of the curve is the same everywhere between these two points. Probabilities are determined by calculating the portion of the rectangle between the two points a and b that is being considered.

In its continuous form, a uniform distribution between a and b has this density function:

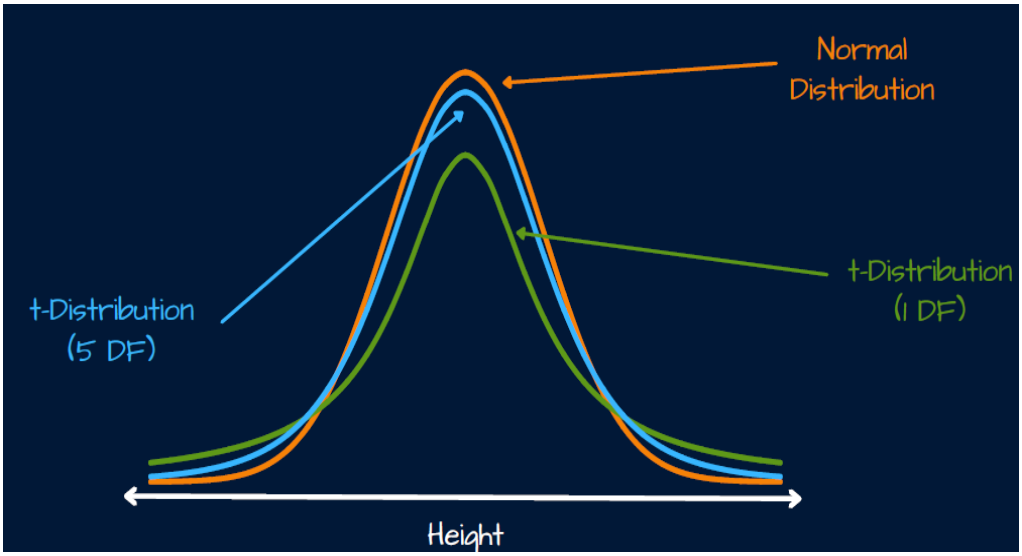
$$f(x) = \frac{1}{b-a} \quad if\ x \in [a,b]\ else\ 0$$

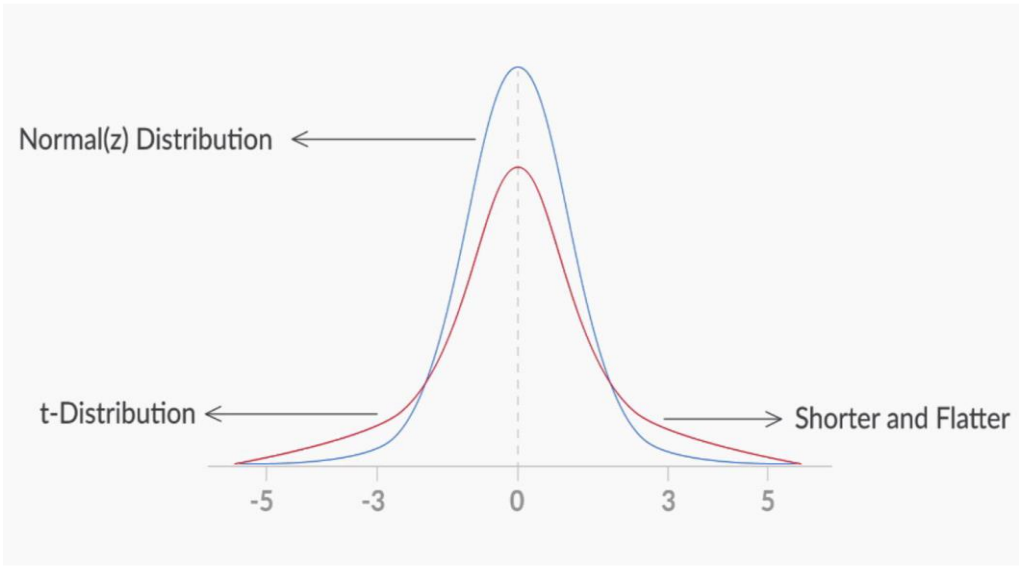
Ex.- The example of uniform distribution is rolling a single dice.here [1,2,3,4,5,6] we have the same probability for each trial.





3.3.3 T-Distribution





The student's t distribution is similar to the normal distribution. The difference is that the tails of the distribution are thicker. This is used when the sample size is small and the population variance is not known. This distribution is defined by the degrees of freedom(p) which is calculated as the sample size minus 1(n – 1).

Each t-distribution is distinguished by what statisticians call degrees of freedom, which are related to the sample size of the data set. If your sample size is n, the degrees of freedom for the

corresponding t-distribution is n -1. For example, if your sample size is 10, you use a t-distribution with 10 -1 or 9 degrees of freedom, denoted t_9. Smaller sample sizes have flatter t-distributions than larger sample sizes. And as you may expect, the larger the sample size is, and the larger the degree of freedom, the more the t-distribution looks like a standard normal distribution or the Z-distribution.

As the sample size increases, degrees of freedom increases the t-distribution approaches the normal distribution and the tails become narrower and the curve gets closer to the mean. This distribution is used to test estimates of the population mean when the sample size is less than 30 and population variance is unknown. The sample variance/standard deviation is used to calculate the t-value.

The PDF is given by,

$$f(t) = \frac{\Gamma\left(\frac{\mathrm{p}+1}{2}\right)}{\sqrt{\mathrm{p}\pi}\,\Gamma\left(\frac{\mathrm{p}}{2}\right)}\left(1+\frac{t^2}{\mathrm{p}}\right)^{-\left(\frac{p+1}{2}\right)}$$

where p is the degrees of freedom and Γ is the gamma function. Check this link for a brief description of the gamma function.

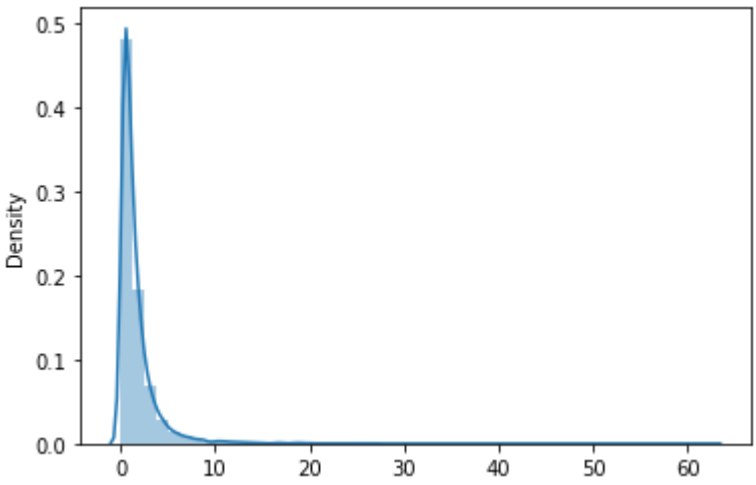
The t-statistic used in hypothesis testing is calculated as follows,

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

Where x̄ is the sample mean, μ the population mean and s is the sample variance

3.3.4 Log-normal Distribution

This distribution is used to plot the random variables whose logarithm values follow a normal distribution. Consider the random variables X and Y. Y = ln(X) is the variable that is represented in this distribution, where ln denotes the natural logarithm of values of X.

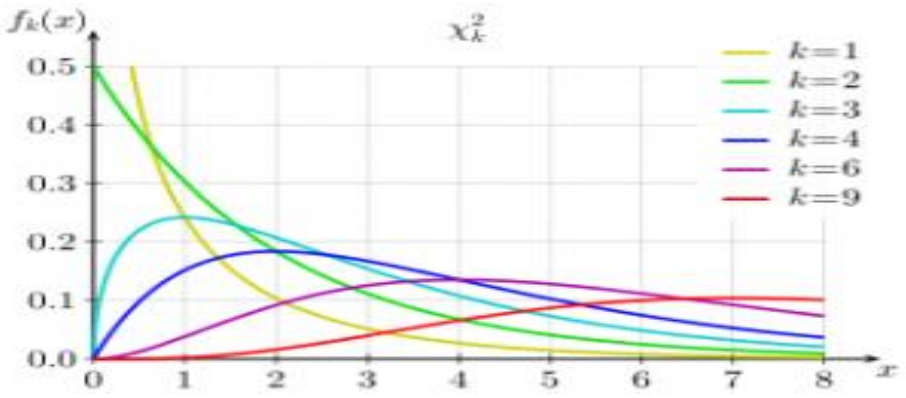


$$f(\mathrm{x}) = \frac{1}{x\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2}\left(\frac{lnx-\mu}{\sigma}\right)^2\right)$$

In the above PDF , μ is the mean of Y and σ is the standard deviation of Y.

3.3.5 Chi-square Distribution

This distribution is equal to the sum of squares of p normal random variables. p is the number of degrees of freedom. Like the t-distribution, as the degrees of freedom increase, the distribution gradually approaches the normal distribution. Below is a chi-square distribution with three degrees of freedom.



The PDF is given by,

$$f(x) = \frac{\left(x^{\frac{p}{2}-1} e^{-\frac{x}{2}}\right)}{2^{p/2}\,\Gamma\left(\frac{p}{2}\right)}$$

where p is the degrees of freedom and Γ is the gamma function.

The chi-square value is calculated as follows:

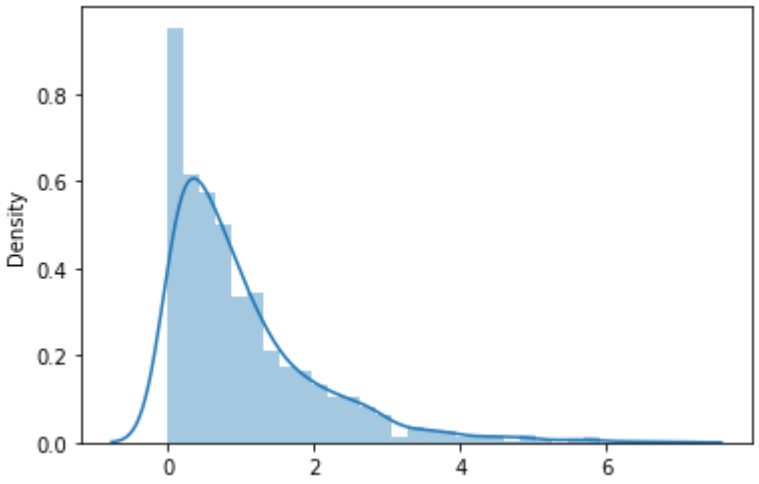
$$\chi^2 = \sum \frac{(o_i - E_i)^2}{E_i}$$

where o is the observed value and E represents the expected value. This is used in hypothesis testing to draw inferences about the population variance of normal distributions.

3.3.6 Exponential Distribution

Recall the discrete probability distribution we have discussed in the Discrete Probability post. In the Poisson distribution, we took the example of calls received by the customer care center. In

that example, we considered the average number of calls per hour. Now, in this distribution, the time between successive calls is explained.



The exponential distribution can be seen as an inverse of the Poisson distribution. The events in consideration are independent of each other.

The PDF is given by,

$$f(\mathrm{x}) = \lambda e^{-\lambda x}$$

where λ is the rate parameter. λ = 1/(average time between events).

CHAPTER 4 Hypothesis Testing

4.1 Understanding Hypothesis Testing

4.1.1 What is a Hypothesis?

When we perform an analysis on a population sample — the analysis could be descriptive, inferential, or exploratory in nature — we get certain information from which we can make claims about the entire population. These are just the claims; we can't be sure if they're actually true. This kind of claim or assumption is called a hypothesis.

Example: The average commute time of employees of a company to and fro office is 35 minutes.

4.1.2 What is Hypothesis Testing?

There are ways to check if your hypothesis has any truth to it, and if the hypothesis is true then apply it to the population parameters. This is called hypothesis testing. The goal is to determine whether there is enough evidence to infer that the hypothesis about the population parameter is true. In hypothesis testing, we confirm our assumptions about the population based on sample data.

4.1.3 Difference between Inferential Statistics & Hypothesis Testing

Inferential statistics is used to find the mean of a population parameter when you have no initial number to start with. So, you start with the sampling activity and find out the sample mean. Then, you estimate the population mean from the sample mean using the confidence interval.

Hypothesis testing is used to confirm your conclusion (or hypothesis) about the population mean (which you know from EDA or your intuition). Through hypothesis testing, you can

determine whether there is enough evidence to conclude if the hypothesis about a population parameter is true or not.

4.1.4 Null & Alternate Hypotheses

Hypothesis Testing starts with the formulation of these two hypotheses:

- **Null hypothesis** (H_o): The status quo
- **Alternate hypothesis** (H_1): The challenge to the status quo

The **null hypothesis** is the prevailing belief about a population; it states that there is no change or no difference in the situation. The **alternate hypothesis**, or **research hypothesis** as it is also called, is the claim that opposes the null hypothesis.

Example: Suppose a man has been charged with murder. In the criminal trial for this case, the jury has to decide whether the defendant is innocent or guilty. Now, this can be turned into two hypotheses. You can claim that the defendant is innocent, and you can claim that the defendant is not innocent, i.e. guilty

Null Hypothesis (H_0) : The defendant is innocent

Here defendant was considered innocent. So, the null hypothesis claims that he is innocent, just like he was before the murder charge.

Alternate hypothesis(H_1) : The defendant is guilty

If you were the prosecutor in the trial, your claim would be that the defendant is guilty, and you would try to prove this.

Note: The hypothesis is always made about the population parameters. The sample parameters are only used as evidence to test the hypothesis.

4.1.5 Outcome of Hypothesis Testing

- If the defendant is found guilty, it means that the jury rejects the null hypothesis in favour of the alternate hypothesis. The jury decides that there is enough evidence to support the alternate hypothesis, and to conclude that the defendant is guilty.
- On the other hand, if the jury acquits the defendant, it means that there is not enough evidence to support the alternate hypothesis. Keep in mind that this does not mean that the defendant is innocent, it just means that there is not enough evidence to conclude that he is guilty. In other words, we cannot accept the null hypothesis; we can only fail to reject it.

Therefore, in hypothesis testing, if there is sufficient evidence to support the alternate hypothesis, you reject the null hypothesis; and if there is not sufficient evidence to support the alternate hypothesis, you fail to reject the null hypothesis. So, you should never say that you "accept" the null hypothesis.

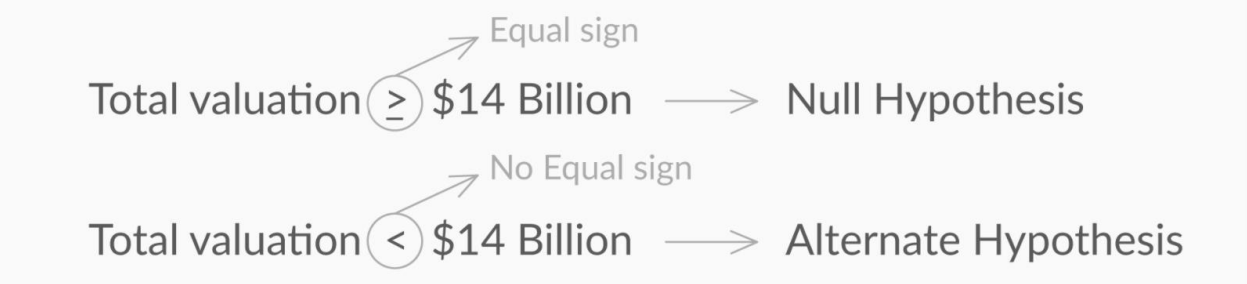
Note: You should never say that you "accept" the null hypothesis.

4.1.6 Formulating Null & Alternate Hypotheses

You can use the following rule to formulate the null and alternate hypotheses:

- **The null hypothesis** always has the following signs: = OR ≤ OR ≥
- **The alternate hypothesis** always has the following signs: ≠ OR > OR <

Situation 1: Flipkart claimed that its total valuation in December 2016 was at least $14 billion. Here, the claim contains ≥ sign (i.e. the at least sign), so **the null hypothesis is the original claim**



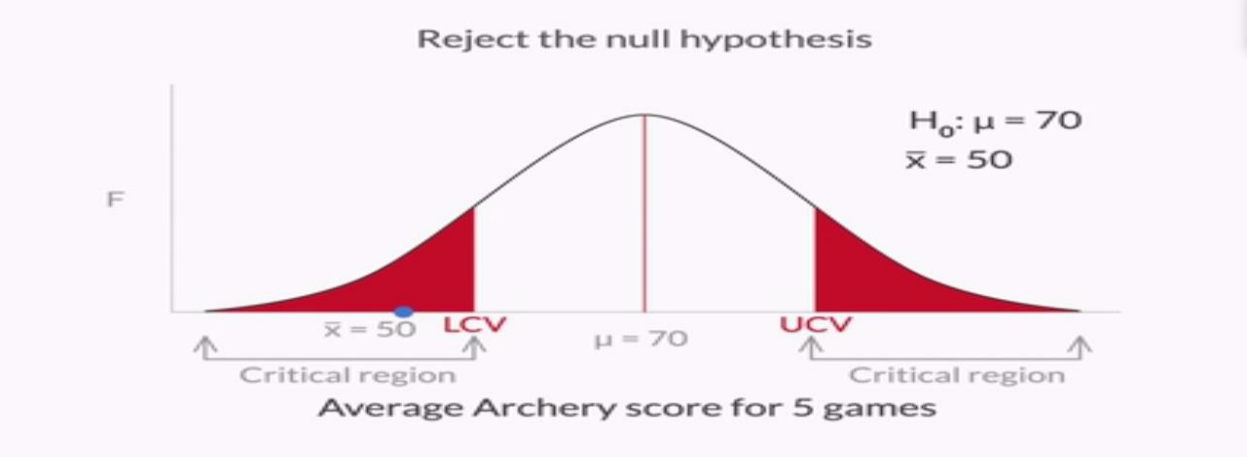
Situation 2: Flipkart claimed that its total valuation in December 2016 was greater than $14 billion. Here, the claim contains > sign (i.e. the 'more than' sign), so **the null hypothesis is the complement of the original claim**.

4.2 Statistical Tests

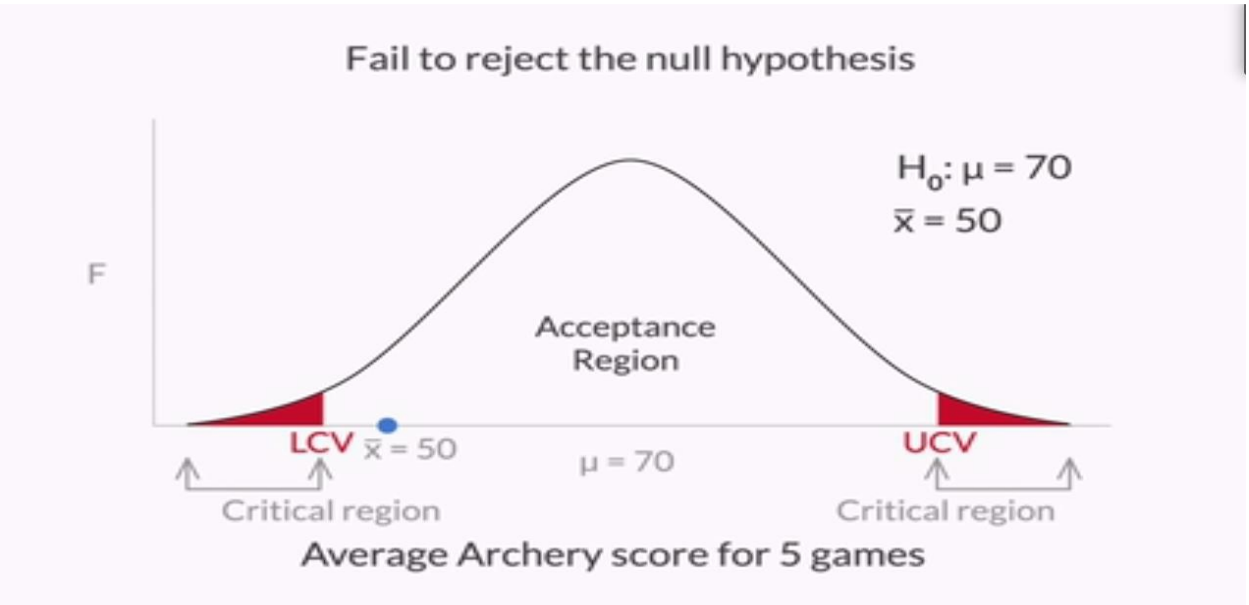
4.2.1 Making Decision

Once you have formulated the null and alternate hypotheses, the next most important step of hypothesis testing is — **making the decision to either reject or fail to reject the null hypothesis**

Situation 1: If sample mean is greater than UCV or less than LCV, i.e. sample mean lies in the criticals region. So reject the null hypothesis



Situation 2: If sample mean is less than UCV or greater than LCV, i.e. sample mean lies in the acceptance region. So we fail to reject the null hypothesis



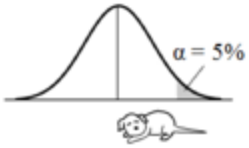
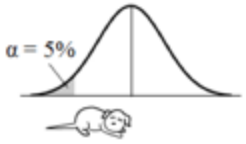
The formulation of the null and alternate hypotheses determines the type of them test and the position of the critical regions in the normal distribution.

You can tell the type of the test and the position of the critical region on the basis of the '**sign' in the alternate hypothesis.**

≠ in H_1 → Two-tailed test → Rejection region on **both sides** of distribution

< in H_1 → Lower-tailed test → Rejection region on **left side** of distribution

> in H_1 → Upper-tailed test → Rejection region on **right side** of distribution

Comparison Operator		Tails of the Test	
H_A	H_0		
≠	=	2-tailed	
>	≤	1- tailed, right-tailed	
<	≥	1-tailed, left-tailed	

4.2.2 Critical Value Method:

After formulating the hypothesis, the steps you have to follow to **make a decision** using **the critical value method** are as follows:

1. Calculate the value of Zċ(Z-Critical) from the given value of α (significance level). Take it a 5% if not specified in the problem.
2. Calculate the critical values (UCV and LCV) from the value of Zċ.

$$Critical\ Value\ (CV) = \mu \pm Z_c \sigma_x$$

Where

$\mu = popilation\ mean\ \ Z_c = Z\ score\ at\ critical\ value\ \sigma_x = sampling\ distribution\ standard\ deviation$

3. Make the decision on the basis of the value of the sample mean x with respect to the critical values (UCV AND LCV).

You can download the z-table from this z-table.

- **α** - **Significance level**(standard error or sampling distribution std deviation) for this test ,it refers to the proportion of the sample mean lying in the critical region. i.e; probability of making an error

  if α=0.05 means 5% of the cases we reject the null hypothesis which is actually true. Other 95% cases we fail to reject the null hypothesis ,when it is true.

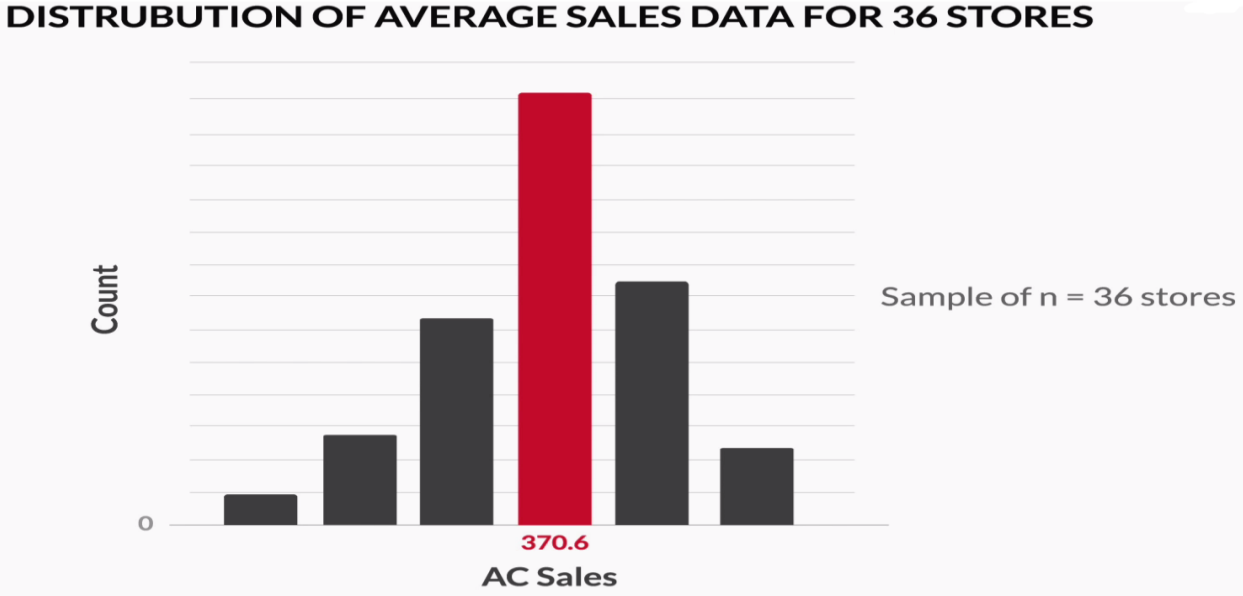
- Sampling distribution is basically the distribution of sample means of a population

Note: You should always examine the evidences with respect to alternative hypothesis NOT with respect to null hypothesis.

- **Example 1:**

Assume that you are the owner of multiple AC stores. You want to know about the mean demand of AC units per month per store during summer. Till now you have been ordering 350 AC units per store per month based on the historic demand. But this time because of intense heat waves, you anticipate that the demand might go up. So you want to check your assumption that the average units required in one month will be different from 350 units per store.

You know that the population standard deviation sigma (σ) is 90, i.e. the distribution obtained every year, containing the sales numbers of every store, has a standard deviation of 90. This year after the sales are over, you take a random sample of 36 stores and plot them. The mean sales turns out to be 370.16. This is your evidence. You can clearly see that it differs from the assumed population mean of 350 units per store.



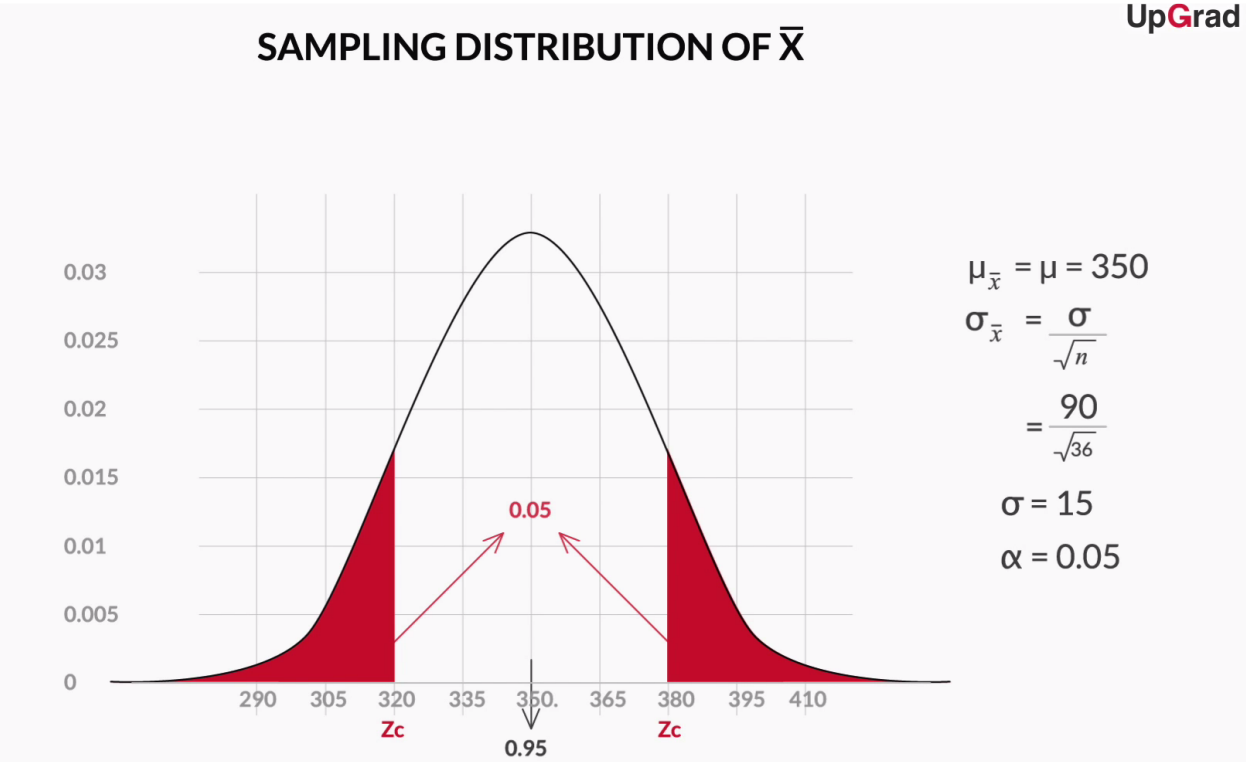


Figure 2 Sampling distribution of sample means

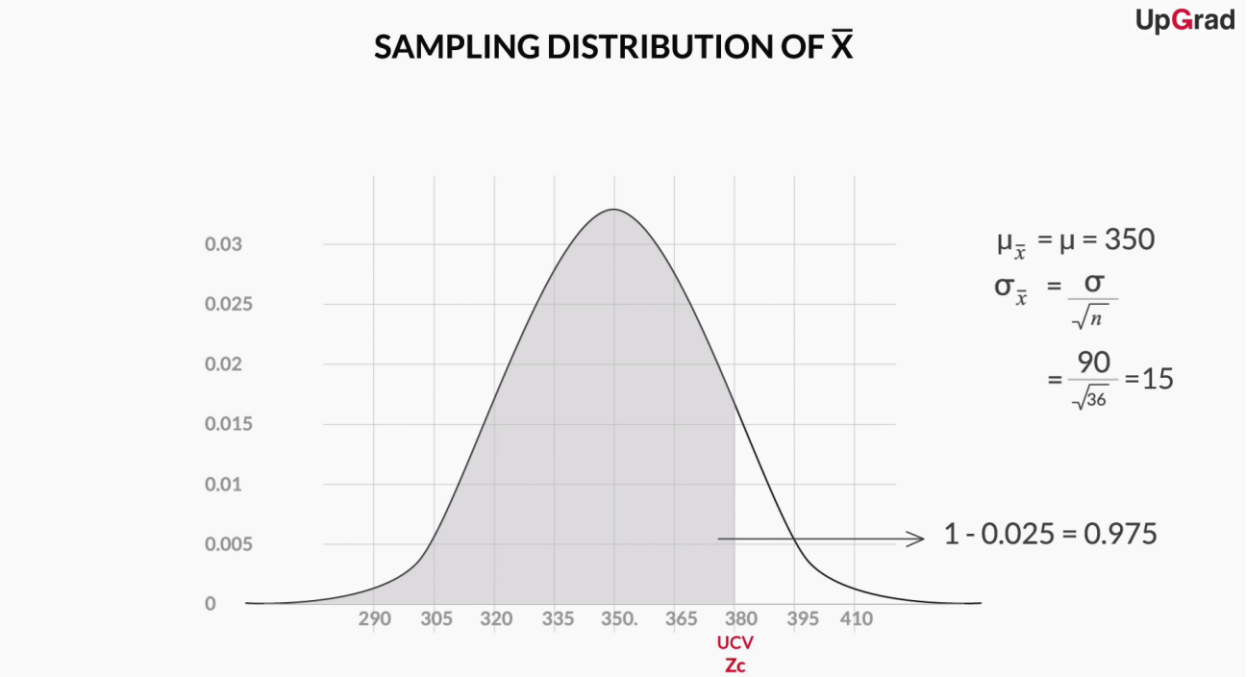
Step 1 : Formulating the hypothesis :

Null Hypothesis Ho : **μ**=350 There is no change in status quo

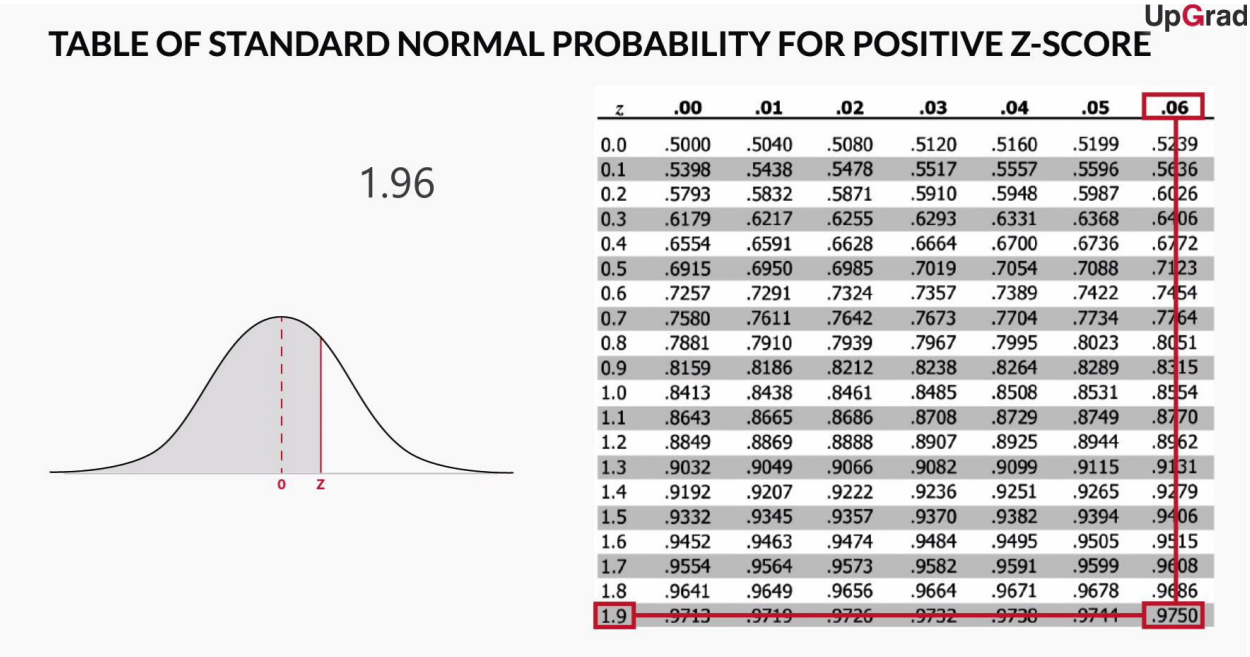
Alternate Hypothesis H1 : $\mu \neq 350$ The status has changed

Step 2: **Making a Decision – Critical Value Method:**

The first step of critical value method is to find Zc. To do that, you calculate the cumulative probability of UCV from the value of α, which is further used to find the z-critical value (Zc) for UCV.

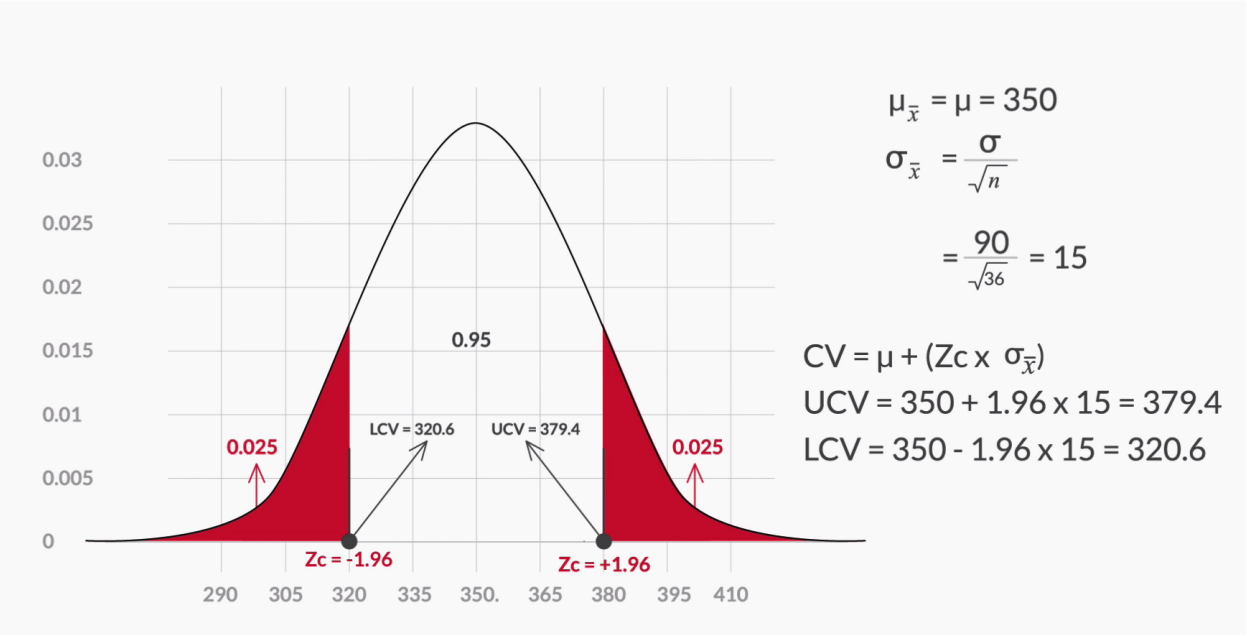


Then you find the z-score of cumulative probability of UCV (Zc in this case).



z	.00	.01	.02	.03	.04	.05	.06
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750

Step 3 :Then you calculate the critical values (UCV and LCV) from the value of Zc.



As sample mean lies is less than UCV and greater than LCV, i.e. it lies in the acceptance region,

Decision: Fail to reject the null hypothesis

- **Example 2 :**

Consider this problem — H_0: $\mu \leq 350$ and H_1: $\mu > 350$

In case of a two-tailed test, you find the z-score of 0.975 in the z-table, since 0.975 was cumulative probability of UCV in that case. In this problem, what would be the cumulative probability of critical point in this example for the same significance level of 5%?

In this problem, the area of the critical region beyond the only critical point, which is on the right side, is 0.05 (in the last problem, it was 0.025). So, the cumulative probability of the critical point (the total area till that point) would be 0.950.

The next step would be to find the Zc, which would basically be the z-score for the value of 0.950. Look at the z-table and find the value of Zc.

0.950 is not there in the z-table. So, look for the numbers nearest to 0.950. You can see that the z-score for 0.9495 is 1.64 (1.6 on the horizontal bar and 0.04 on the vertical bar), and the z-score for 0.9505 is 1.65. So, taking the average of these two, the z-score for 0.9500 is 1.645.

So, the Zc comes out to be 1.645. Now, find the critical value for the given Zc and make the decision to accept or reject the null hypothesis.

$\mu = 350$ $\sigma = 90$ N (Sample size) = 36 $\bar{x} = 370.16$

The critical value can be calculated from $\mu + Zc \times (\sigma/\sqrt{N})$. $350 + 1.645(90/\sqrt{36}) = 374.67$. Since 370.16 ($\bar{x}$) is less than 374.67, $\bar{x}$ lies in the acceptance region and you fail to reject the null hypothesis.

- **Example 3:**

Government regulatory bodies have specified that the maximum permissible amount of lead in any food product is 2.5 parts per million or 2.5 ppm. Let's say you are an analyst working at the food regulatory body of India FSSAI. Suppose you take 100

random samples of Sunshine from the market and have them tested for the amount of lead. The mean lead content turns out to be 2.6 ppm with a standard deviation of 0.6.

One thing you can notice here is that the standard deviation of the sample is given as 0.6, instead of the population's standard deviation. In such a case, you can approximate the population's standard deviation to the sample's standard deviation, which is 0.6 in this case.

Answer the following questions in order to find out if a regulatory alarm should be raised against Sunshine or not, at 3% significance level.

H_0: Average lead content ≤ 2.5 ppm and H_1: Average lead content > 2.5 ppm
✓ **Correct**

Feedback:
The null hypothesis is your assumption about the population — it is based on the status quo. It always makes an argument about the population using the equality sign. The null hypothesis in this case would be that the average lead content in the food material is less than or equal to 2.5 ppm. And the alternate hypothesis is that the average lead content is greater than 2.5 ppm

Calculate the z-critical score for this test at 3% significance level.

This is a one-tailed test. So, for 3% significance level, you would have only one critical region on the right side with a total area of 0.03. This means that the area till the critical point (the cumulative probability of that point) would be 1 - 0.030 = 0.970. So, you need to find the z-value of 0.970. The z-score for 0.9699 (~0.970) in the z-table is 1.88.

Now, you need to find out the critical values and make a decision on whether to raise a regulatory alarm against Sunshine or not. Select the correct option.

The critical value can be calculated from μ + Zc x (σ/√N), as 2.5 + 1.88(0.6 /√100) = 2.61 ppm . You need to use the + sign since the critical value is on the right-hand side

(upper-tailed test). Since the sample mean 2.6 ppm is less than the critical value (2.61 ppm), you fail to reject the null hypothesis and don't raise a regulatory alarm against Sunshine.

The critical value for this test at 3% significance level comes out to be 2.61 ppm. If you take more than 100 samples (with the same sample mean and standard deviation), how would the z-score and critical value change?

The z-score would remain the same but the critical value would decrease

Since Zc is calculated from the given value of α (3%), it remains the same. Critical value is calculated using the formula: μ + Zc x (σ/√N), since it is an upper-tailed test. If you increase the value of N, the critical value would decrease according to the formula.

What exactly is the value of σ/√n?

This part might seem a bit confusing if you aren't thorough with the concepts of Sampling distribution from the Inferential Statistics module. This value of σ/√n isn't the standard deviation of the sample or the population but rather it is the standard deviation of the sampling distribution for the given sample. We conduct our hypothesis test on the sampling distribution only. Thus when the standard deviation of the population isn't given, we take the sample's standard deviation to calculate the sampling distribution's standard deviation. This is a very important concept to know and you should be pretty confident in how we are manipulating the values here. We used this same concept in calculating the confidence intervals. In that case we used S/√n instead which denoted that we can only use the sample's standard deviation in that case.

- **Example 4:**

The water purifier company Kent claims that the total hardness of the water after being treated and filtered by its product Kent RO is less than 300 ppm on an average. To test the claim, a water inspector takes a sample of 400 purifiers and calculates the

mean total hardness of water being filtered out, which comes out to be 296 ppm, with a standard deviation of 25 ppm. The hypothesis test is to be conducted at a significance level of 3%.

H_0: $\mu \geq 300$ ppm and H_1: $\mu < 300$ ppm; Critical region lies on the left side of the tail

Observe that the claim statement has a less than sign associated with it. Thus the null hypothesis would be the complement of it. And the alternate hypothesis would be the claim statement itself. Once you formulate the null and alternate hypothesis,i.e. H_0: $\mu \geq 300$ ppm and H_1: $\mu < 300$ ppm, it is easy to see that the critical region would lie on the left side of the tail since the alternate hypothesis has a < sign associated with it.

What would be the Zc for this case?

From the significance level = 3 % we get the area of the critical region to be 0.03. Now this region would lie on the left side of the tail. Thus we have P(Z<Zc)= 0.03. From the table, we get the value of Zc = -1.88

what would be the critical value and the final decision to be made here?

297.65; Reject the null hypothesis

You need to calculate the critical value from the Zc calculated in the previous question. Here you get the value of Zc as -1.88. Now critical value is 300 - 1.88*1.25 = 297.65 runs. Observe that 296 falls in the critical region. Hence you need to reject the null hypothesis in this case.

4.2.3 p-value Method

There are various methods similar to the critical value method to statistically make your decision about the hypothesis. One such method is the p-value method. This is an important method and is used more frequently in the industry.

What is p-value?

A P-value measures the strength of evidence in support of a null hypothesis. Suppose the test statistic in a hypothesis test is equal to *K*. The P-value is the probability of observing a test statistic as extreme as *K*, assuming the null hypothesis is true. If the P-value is less than the significance level, we reject the null hypothesis.

If p =0.8 means that probability of not rejecting the null hypothesis is 80%, so we fail to reject the null hypothesis . If p<0.05 then probability of evidence supporting null hypothesis is very less , so we can reject it.

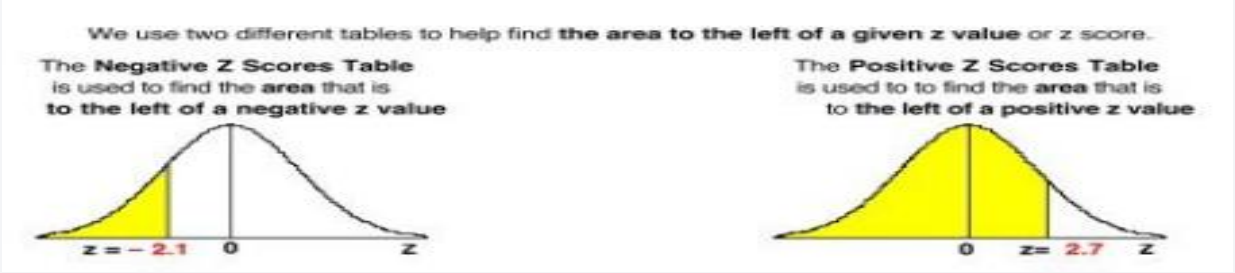
Note1: In simple words, it is the **probability that the null hypothesis will not be rejected.**

Note 2: Generally p=0.05 is considered as significance level (α) or standard error or confidence level of 5%.

After formulating the null and alternate hypotheses, the steps to follow in order to **make a decision** using the **p-value method** are as follows:

1. Calculate the value of the z-score for the sample mean point on the distribution.
2. Calculate the p-value from the cumulative probability for the given z-score using the z-table.
3. Make a decision on the basis of the p-value (multiply it by 2 for a two-tailed test) with respect to the given value of α (significance value).

To find the correct p-value from the z-score, find the **cumulative probability** first, by simply looking at the z-table, which gives you the area under the curve till that point.



Situation 1: The sample mean is on the left side of the distribution mean (the z-score is negative).

Example: The z-score for the sample point = -3.02

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026

Cumulative probability of the sample point = 0.0013

For a one-tailed test: p = 0.0013

For a two-tailed test: p = 2 * 0.0013 = 0.0026

Situation 2: The sample mean is on the right side of the distribution mean (the z-score is positive).

Example: z-score for sample point = + 3.02

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Cumulative probability of the sample point = 0.9987

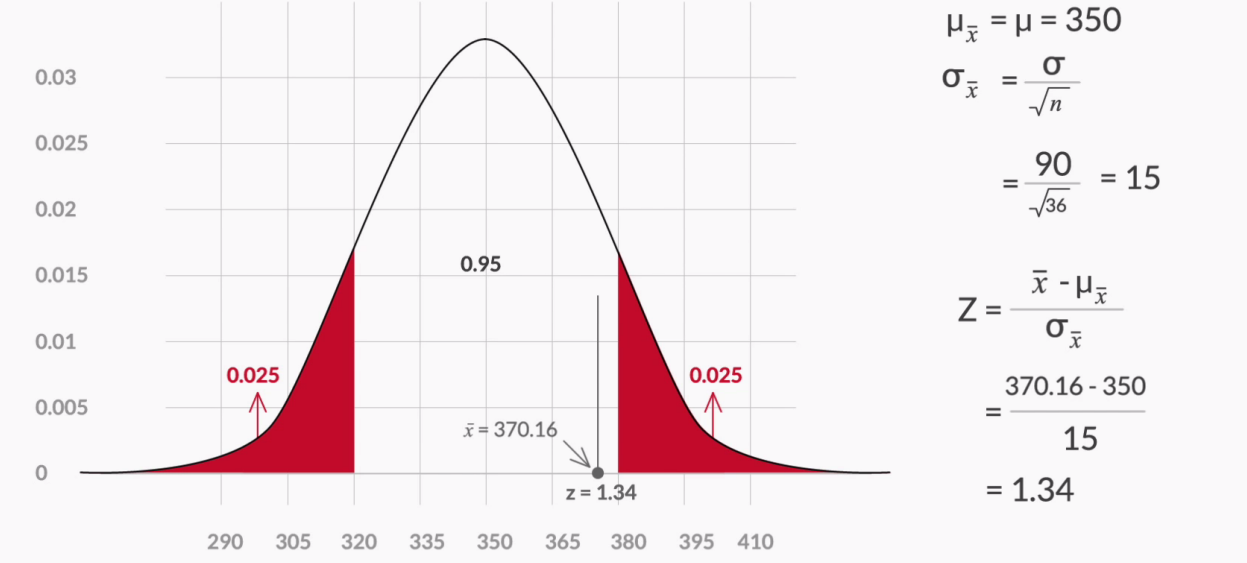
For a one-tailed test: p = 1 - 0.9987 = 0.0013

For a two-tailed test: p = 2 (1 - 0.9987) = 2 * 0.0013 = 0.0026

- **Example1:**

Taking the same AC sales problem, hypotheses for the situation would remain the same.

Step1: So, you start by finding out the z-value for given sample mean.



$$\mu_{\bar{x}} = \mu = 350$$

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

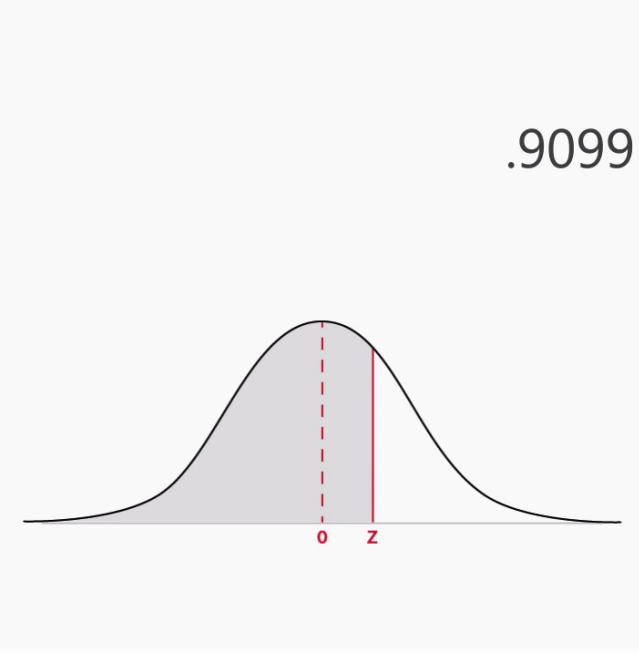
$$= \frac{90}{\sqrt{36}} = 15$$

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

$$= \frac{370.16 - 350}{15}$$

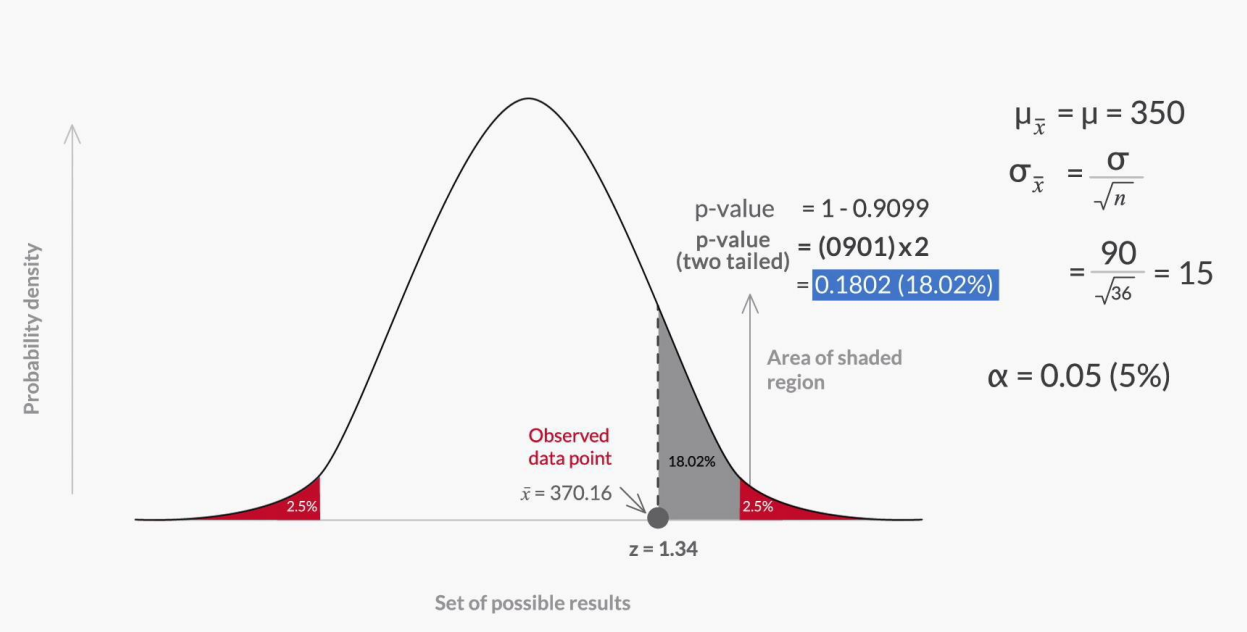
$$= 1.34$$

Step2 : you find the cumulative propability for the calculated z-value of sample mean.



z	.00	.01	.02	.03	.04
0.0	.5000	.5040	.5080	.5120	.5160
0.1	.5398	.5438	.5478	.5517	.5557
0.2	.5793	.5832	.5871	.5910	.5948
0.3	.6179	.6217	.6255	.6293	.6331
0.4	.6554	.6591	.6628	.6664	.6700
0.5	.6915	.6950	.6985	.7019	.7054
0.6	.7257	.7291	.7324	.7357	.7389
0.7	.7580	.7611	.7642	.7673	.7704
0.8	.7881	.7910	.7939	.7967	.7995
0.9	.8159	.8186	.8212	.8238	.8264
1.0	.8413	.8438	.8461	.8485	.8508
1.1	.8643	.8665	.8686	.8708	.8729
1.2	.8849	.8869	.8888	.8907	.8925
1.3	.9032	.9049	.9066	.9082	.9099
1.4	.9192	.9207	.9222	.9236	.9251
1.5	.9332	.9345	.9357	.9370	.9382
1.6	.9452	.9463	.9474	.9484	.9495
1.7	.9554	.9564	.9573	.9582	.9591
1.8	.9641	.9649	.9656	.9664	.9671
1.9	.9713	.9719	.9726	.9732	.9738
2.0	.9772	.9778	.9783	.9788	.9793
2.1	.9821	.9826	.9830	.9834	.9838
2.2	.9861	.9864	.9868	.9871	.9875
2.3	.9893	.9896	.9898	.9901	.9904
2.4	.9918	.9920	.9922	.9925	.9927
2.5	.9938	.9940	.9941	.9943	.9945
2.6	.9953	.9955	.9956	.9957	.9959

Then you find the p-value by the approach mentioned above.



Finding p-value from cumulative probability
As p-value (0.1802) is greater than the value of α (0.05),
Decision: Fail to reject the null hypothesis

- **Example 2:**

Let's say you work at a pharmaceutical company that manufactures an antipyretic drug in tablet form, with paracetamol as the active ingredient. An antipyretic drug reduces fever. The amount of paracetamol deemed safe by the drug regulatory authorities is 500 mg. If the value of paracetamol is too low, it will make the drug ineffective and become a quality issue for your company. On the other hand, a value that is too high would become a serious regulatory issue.

There are 10 identical manufacturing lines in the pharma plant, each of which produces approximately 10,000 tablets per hour.

Your task is to take a few samples, measure the amount of paracetamol in them, and test the hypothesis that the manufacturing process is running successfully, i.e., the paracetamol content

is within regulation. You have the time and resources to take about 900 sample tablets and measure the paracetamol content in each.

Upon sampling 900 tablets, you get an average content of 510 mg with a standard deviation of 110. What does the test suggest if you set the significance level at 5%? Should you be happy with the manufacturing process, or should you ask the production team to alter the process? Is it a regulatory alarm or a quality issue?

One thing you can notice here is that the standard deviation of the sample of 900 is given as 110 instead of the population standard deviation. In such a case, you can **assume the population standard deviation to be the same as the sample standard deviation, which is 110 in this case**.

Null hypothesis: $\mu=500$

Alternate Hypothesis: $\mu\neq500$

Calculate the Z-score for the sample mean ($\bar{x}$) = 510 mg.

You can calculate the Z-score for the sample mean of 510 mg using the formula: $(\bar{x} - \mu) / (\sigma / \sqrt{N})$. This gives you $(510 - 500)/(110/\sqrt{900}) = (10)/(110/30) = 2.73$. Notice that since the sample mean lies on the right side of the hypothesised mean of 500 mg, the Z-score comes out to be positive

Find out the p-value for the Z-score of 2.73 (corresponding to the sample mean of 510 mg)

The value in the Z-table corresponding to 2.7 on the vertical axis and 0.03 on the horizontal axis is 0.9968. Since the sample mean is on the right side of the distribution

and this is a two-tailed test (because we want to test whether the value of the paracetamol is too low or too high), the p-value would be 2 * (1 - 0.9968) = 2 * 0.0032 = 0.0064

Based on this hypothesis test, what decision would you make about the manufacturing process?

Here, the p-value comes out to be 0.0064. Here, the p-value is less than the significance level (0.0064 < 0.05) and a smaller p-value gives you greater evidence against the null hypothesis. So, you reject the null hypothesis that the average amount of paracetamol in medicines is 500 mg. So, this is a regulatory alarm for the company, and the manufacturing process needs to change.

- **Example 3:**

A nationwide survey claimed that the unemployment rate of a country is at least 8%. However, the government claimed that the survey was wrong and the unemployment rate is less than that. The government asked about 36 people, and the unemployment rate came out to be 7%. The population standard deviation is 3%

Null Hypothesis $H0:\mu \geq 8\%$

Aleternate Hypothesis $H1:\mu < 8\%$

Based on the information above, conduct a hypothesis test at a 5% significance level using the p-value method. What is the Z-score of the sample mean point $\bar{x} = 7\%$.

$\mu=8\%;\sigma=3\%;n=36;\bar{x}=7\%;S.E.=3/\sqrt{36}=0.5$. Now, $Z\bar{x}=(7-8)/0.5=-2$

Calculate the p-value from the cumulative probability for the given Z-score using the Z-table. In other words, find out the p-value for the Z-score of -2.0 (corresponding to the sample mean of 7%).

The p-value corresponding to a Z-score of -2.0 is 0.0228

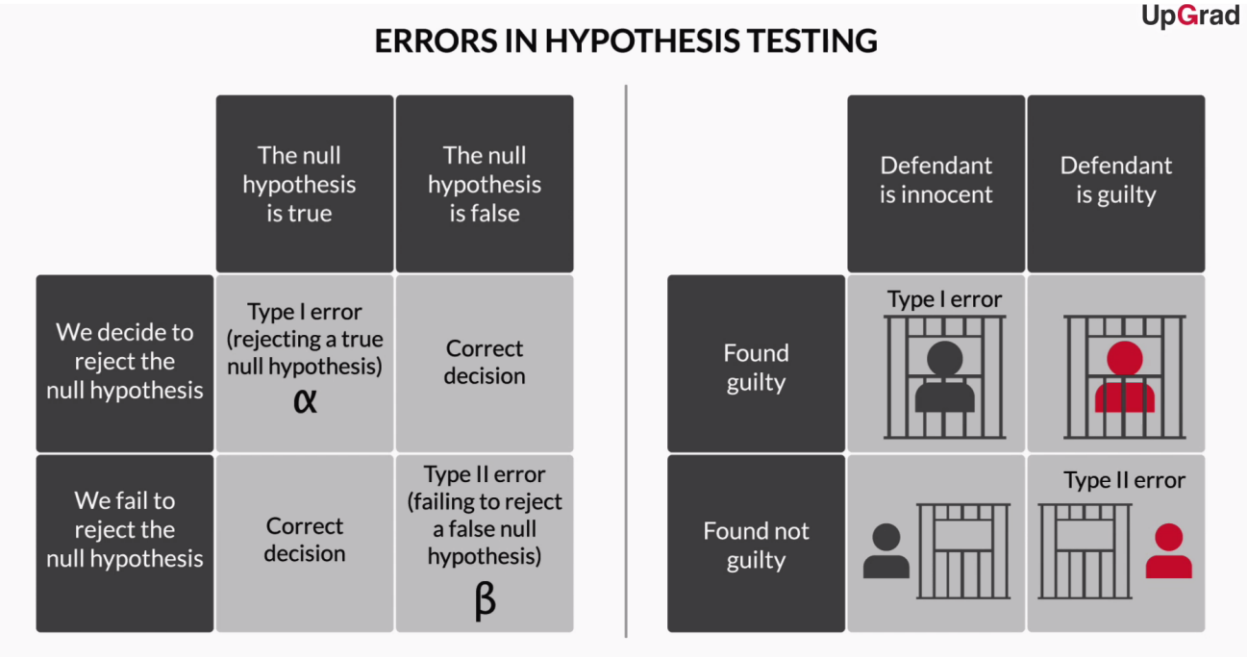
Make the decision on the basis of the p-value with respect to the given value of α (significance value)

You reject the null hypothesis because the p-value is less than 0.05.

4.2.4 Types of errors

There are two possible errors we can commit during hypothesis testing

- The **type I error** occurs when the null hypothesis is true(reality is true) but we reject it, i.e. reject H0 when it is true.(FalsePositive)
- The **type II error** occurs when the null hypothesis is false (reality is false) but we fail to reject it, i.e. fail to reject H0 when it is false.(False Negative)



In a type-1 error, the Ho is correct, but it is rejected

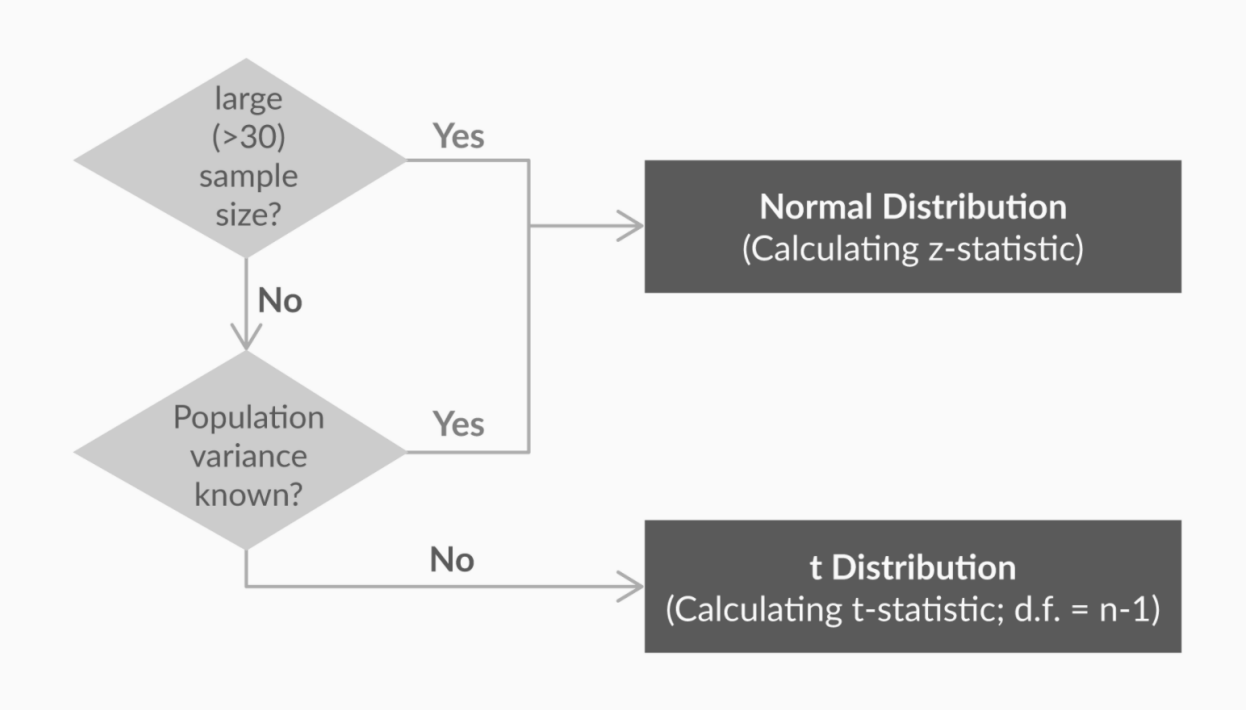
In a type-2 error, the H_0 is incorrect. but fail to reject the null hypothesis(Ho is accepted).

Type I and Type II Error

Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = 1 - β
Not rejected	Correct decision True negative Probability = 1 - α	Type II error False negative Probability = β

4.2.5 T-test

The most important use of the t-distribution is that you can approximate the value of the **standard deviation of the population (σ)** from the **sample standard deviation (s)**. However, as the sample size increases more than 30, the t-value tends to be equal to the z-value.



If you look at how the method of **making a decision** changes if you are using the sample's standard deviation instead of the population's. If you recall the critical value method, the first step is as follows:
1. Calculate the value of Zc from the given value of α (significance level). Take it as 5% if not specified in the problem.

So, to find Zc, you would use the **t-table** instead of the z-table. The **t-table** contains values of Zc for a given degree of freedom and value of α (significance level). Zc, in this case, can also be called as t-statistic (critical).

The **degrees of freedom**, *k*, are the number of values that are free to vary. i.e., one less than the number of observations.

Example: Say you have 4 numbers that add up to 1: i.e., A+B+C+D=1

How many of the values are free to vary?

The answer is 3 because if you know 3 of the numbers, then you can solve for the 4th one:

D=1−(A+B+C)

So, this example has 3 **degrees of freedom.**

- **Example:**

According to a study, the daily average time spent by a user on a social media website is 50 minutes. To test the claim of this study, Ramesh, a researcher, takes a sample of 25 website users and finds out that the mean time spent by the sample users is 60 minutes and the sample standard deviation is 30 minutes.
Based on this information, the null and the alternative hypotheses will be:
Ho = The average time spent by the users is 50 minutes
H1= The average time spent by the users is not 50 minutes
Use a 5% significance level to test this hypothesis. Use the t-table from the link here.

Here, the population standard deviation is not known and the sample size is less than 30; so, we can use the t-test.
Therefore, the t-test statistic (t) = $(x - \mu)/(s/\sqrt{(n)}) = (60 - 50)/(30/\sqrt{25}) = 1.66$.
The sample size is 25. So, the degree of freedom = 25 - 1 = 24.
This is a two-tailed t-test with a significance level of 5%; so, the critical t-value = t 0.05, 24 = 2.064 (using the t-table).
Since the observed t-value is smaller than the critical value, Ramesh fails to reject the null hypothesis.

4.2.6 Paired t-test

🔗 As data scientists, we constantly strive to improve our machine learning models to achieve the best possible performance. However, it's not always clear which model is truly better. In some cases, the difference in performance between the two models can be small, and it may be difficult to determine if it's just due to random chance or if it's statistically significant. That's where the statistical significance tests such as paired t-test come in.

❓ The paired t-test is a statistical test used to compare the performance of two models on the same dataset. It evaluates whether there is a significant difference in the mean of their

performance metric, such as accuracy or mean squared error. The paired t-test is a powerful tool in evaluating machine learning models because it considers the data's variability, allowing us to determine whether the observed difference is due to chance or to a real difference in performance.

Here's the formula for the paired t-test:

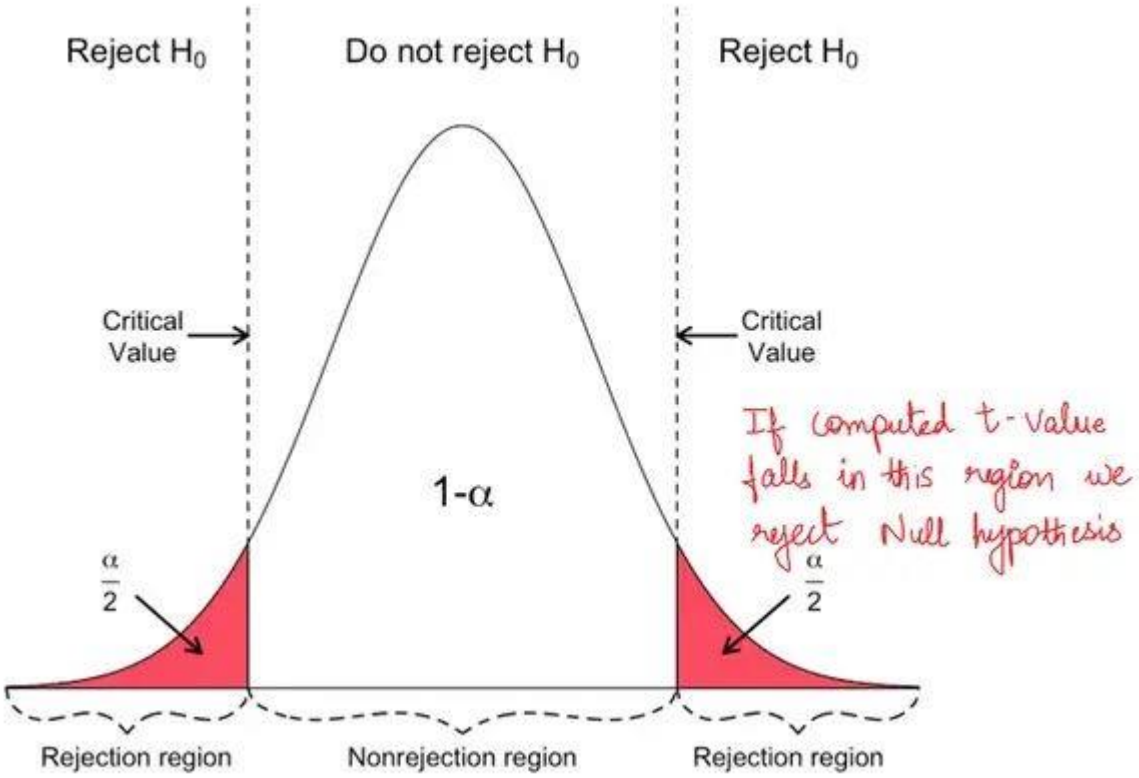
t = (mean of the differences) / (standard deviation of the differences/sqrt (n))

t is the t-statistic, which measures the difference between the means of the two models relative to the variability of the data

If the calculated t-statistic is greater than the critical value at the desired level of significance, then we can reject the null hypothesis that the means are equal, and conclude that there is a statistically significant difference in performance between the two models.

The paired t-test is a powerful tool in evaluating machine learning models, but it's essential to use it correctly. One key assumption of the paired t-test is that the differences in performance metric between the two models are normally distributed. Additionally, the data should be paired meaningfully, such as using the same training and test datasets for both models.

👀 In summary, the paired t-test, by comparing the performance of two models on the same dataset, allows us to determine whether the observed difference is due to chance or to a real difference in performance. However, it's important to use the test correctly and to ensure that the data meets the necessary assumptions



4.2.7 Two sample tests:

Note: tests for comparing a normally distributed set of measurements with a hypothesized value (one-sample) and for comparing the means between two groups (two-sample test).

Two sample mean test-paired	paired is used when your **sample observations are from the same individual** or object. During this test, you are testing the same subject twice. **For example**, if you are testing a new drug, you would need to compare the sample before and after the drug is taken to see if the results are different.
Two-sample mean test - unpaired	**is used when your sample observations are independent**. During this test, you are not testing the same subject twice. **For example**, if you are testing a new drug, you would compare its effectiveness to

	that of the standard available drug. So, you would take a sample of patients who consumed the new drug and compare it with those who consumed the standard drug
Two-sample proportion test	is used **when your sample observations are categorical**, **with two categories**. It could be True/False, 1/0, Yes/No, Male/Female, Success/Failure etc. For example, if you are comparing the effectiveness of two drugs, you would define the desired outcome of the drug as the success. So, you would take a sample of patients who consumed the new drug and record the number of successes and compare it with successes in another sample who consumed the standard drug
A/B testing	is a **direct industry application of the two-sample proportion test sample**. While developing an e-commerce website, there could be different opinions about the choices of various elements, such as the shape of buttons, the text on the call-to-action buttons, the colour of various UI elements, the copy on the website, or numerous other such things. Often, the choice of these elements is very subjective and is difficult to predict which option would perform better. To resolve such conflicts, you can use A/B testing. **A/B testing** provides a way for you to test two different versions of the same element and see which one performs better. A/B testing is entirely based on the two-sample proportion test, as the two-sample proportion test is used when you want to compare the proportions of two different samples. You can use various tools to conduct A/B testing (or two-sample proportion test) like R, Optimizely etc

4.2.8 A/B testing

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

How does A/B Testing Work?

In this section, let's understand through an example the logic and methodology behind the concept of A/B testing.

Let's say there is an e-commerce company XYZ. It wants to make some changes in its newsletter format to increase the traffic on its website. It takes the original newsletter and marks it A and makes some changes in the language of A and calls it B. Both newsletters are otherwise, the same in color, headlines, and format.

Objective

Our objective here is to check which newsletter brings higher traffic on the website i.e the conversion rate. We will use A/B testing and collect data to analyze which newsletter performs better.

Step 1: Make a Hypothesis

- Null hypothesis or H0:

" there is no difference in the conversion rate in customers receiving newsletter A and B".

- Alternative Hypothesis or H0:

Ha is- "the conversion rate of newsletter B is higher than those who receive newsletter A"

Step 2:. Create Control Group and Test Group

The Control Group is the one that will receive newsletter A and the Test Group is the one that will receive newsletter B.

For this experiment, we randomly select 1000 customers – 500 each for our Control group and Test group.

Step 3: Conduct the A/B Test and Collect the Data

One way to perform the test is to calculate daily conversion rates for both the treatment and the control groups. Since the conversion rate in a group on a certain day represents a single data point, the sample size is actually the number of days. Thus, we will be testing the difference between the mean of daily conversion rates in each group across the testing period. When we ran our experiment for one month, we noticed that the mean conversion rate for the Control group is 16% whereas that for the test Group is 19%.

Step 4: Statistical significance of test

The difference between your control version and the test version is not due to some error or random chance. To prove the statistical significance of our experiment we can use a two-sample T-test.

The two–sample t–test is one of the most commonly used hypothesis tests. It is applied to compare the average difference between the two groups.

Step 5 : Result

For our example, the observed value i.e the mean of the test group is 0.19. The hypothesized value (Mean of the control group) is 0.16. On the calculation of the t-score, we get the t-score as 3.787. and the p-value is 0.00036.

So what does all this mean for our A/B Testing?

Here, our p-value is less than(0.00036 < 0.05) the significance level i.e 0.05. Hence, we can reject the null hypothesis. This means that in our A/B testing, newsletter B is performing better than newsletter A. So our recommendation would be to replace our current newsletter with B to bring more traffic to our website.

What Mistakes Should We Avoid While Conducting A/B Testing?

There are a few key mistakes I've seen data science professionals making. Let me clarify them for you here:

● **Invalid hypothesis**: The whole experiment depends on one thing i.e the hypothesis. What should be changed? Why should it be changed, what the expected outcome is, and so on? If you start with the wrong hypothesis, the probability of the test succeeding, decreases

● **Testing too Many Elements Together**: Industry experts caution against running too many

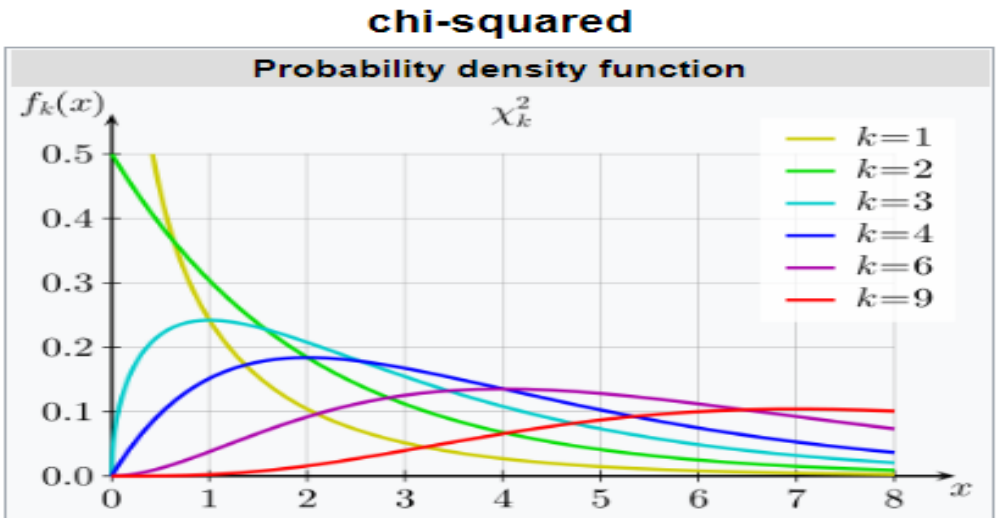
tests at the same time. Testing too many elements together makes it difficult to pinpoint which element influenced the success or failure. Thus, prioritization of tests is indispensable for successful A/B testing

● **Ignoring Statistical Significance**: It doesn't matter what you feel about the test. Irrespective of everything, whether the test succeeds or fails, allow it to run through its entire course so that it reaches its statistical significance

● **Not considering the external factor**: Tests should be run in comparable periods to produce meaningful results. For example, it is unfair to compare website traffic on the days when it gets the highest traffic to the days when it witnesses the lowest traffic because of external factors such as sale or holidays

4.2.9 Chi-Square test (Source)

It is a one-tail test. A Pearson's **chi-square test** (kai-square) is a statistical test for categorical data. It is used to determine whether your data are significantly different from what you expected. 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 single categorical variable is different from your expectations. (This is used to test whether the sample data correctly represents the population data.) determines if sample data matches a population. To test if the sample is coming from a population with specific distribution.
- The **chi-square test of independence** is used to test whether two categorical variables are related to each other.
- 

- **Chi-squared test of independence**:

This is used to determine whether or not there is a significant relationship between two nominal (categorical) variables.

For example, a researcher wants to examine the relationship between gender (male vs female) and the chances of developing Alzheimer's disease. The chi-squared test of independence can be used to examine this relationship. The null hypothesis (Ho) for this test is that there is no relationship between gender and life expectancy, and the alternative hypothesis is that there is a relationship between gender and life expectancy.

Here, there are two categorical variables (nominal variables): male and female.
The expected value is calculated by assuming that the null hypothesis is correct. So, if you select a sample of, say, 100 Alzheimer's patients, 50 should be men and 50 should be women. Let's say the sample value comes out to be a bit different, and in a sample of 100 Alzheimer's patients, 60 are men and 40 are women.

	Male	**Female**
Expected Value (E)	50	50
Sample Value(Observed value) (O)	60	40

The test statistic for the chi-squared test is equal to $\boldsymbol{\chi^2 = \Sigma(O - E)^2/E}$
where O is the observed sample value and E is the expected value.
So, our test statistic will be equal to:
χ2 = (10^2)/50 + (10^2)/50 = 4

Let's select the level of significance as 5%, or 0.05.

Degrees of freedom = (r - 1) x (c - 1), where r is the number of rows and c is the number of columns.

So, the degree of freedom, in this case, is 1.

Now, you will use the chi-squared distribution table to calculate the critical value.

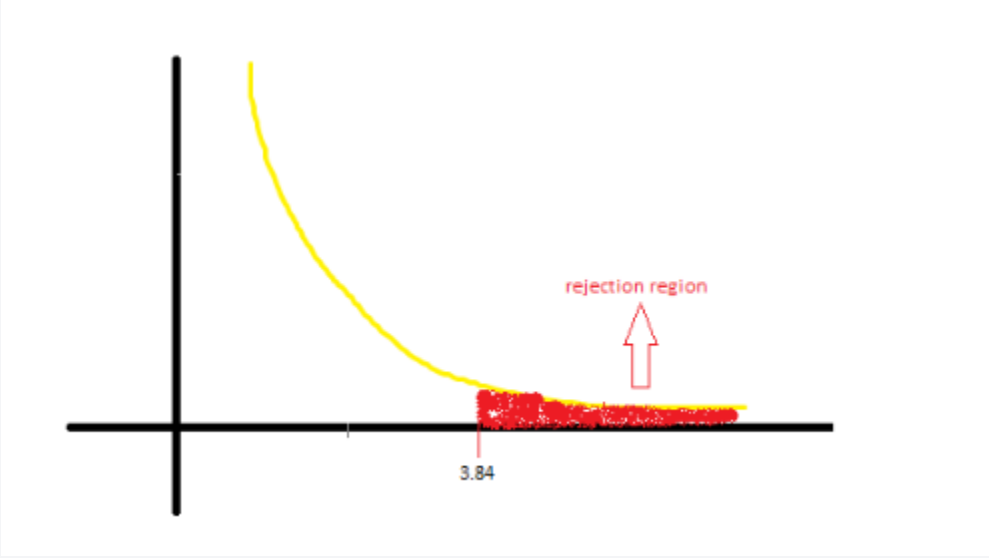
Select the value corresponding to the required degrees of freedom and the significance level

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49
34	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06
38	19.29	20.69	22.88	24.88	27.34	49.51	53.38	56.90	61.16
42	22.14	23.65	26.00	28.14	30.77	54.09	58.12	61.78	66.21
46	25.04	26.66	29.16	31.44	34.22	58.64	62.83	66.62	71.20
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15
55	31.73	33.57	36.40	38.96	42.06	68.80	73.31	77.38	82.29
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38

So, the critical value is 3.84, and the test statistic value is 4

In this case, the test statistic value (4), which is greater than the critical value, lies in the rejection region. Therefore, you reject the null hypothesis.



Example 2:

Example: A teacher wants to know the answer to whether the outcome of a mathematics test is related to the gender of the person taking the test.

Step 1: Calculate the row and column total of the above contingency table:

	Boys	Girls	Total
Pass	17	20	37
Fail	8	5	13
Total	25	25	50

Step 2: Calculate the expected frequency for each individual cell by multiplying row sum by column sum and dividing by total number:

Expected Frequency = (Row Total x Column Total)/Grand Total

For the first cell, the expected frequency would be (37*25)/50 = 18.5.

For the second cell, the expected frequency would be (13*25)/50= 6.5

Step 3: Calculate the value of chi-square using the formula:

Calculate the right-hand side part of each cell. For example, for the first cell, ((17-18.5)^2)/18.5 =0.1216.

Step 4: Then, add all the values obtained for each cell. In this case, the values are:

0.1216+0.1216+0.3461+0.3461 = 0.9354

Step 5: Calculate the

degrees of freedom = (Number of rows-1)*(Number of columns-1)

= 1*1 = 1

The next task is to compare it with the critical chi-square value from the table we saw above.

The Chi-Square calculated value is 0.9354 which is less than the critical value of 3.84.

So in this case, we **fail to reject the null hypothesis**. This means there is no significant association between the two variables, i.e, boys and girls have a statistically similar pattern of pass/fail rates on their mathematics tests.

- **Chi-Square goodness of fit:**

If you have a single measurement variable, you use a Chi-Square goodness of fit.
Example: A coin is flipped 100 times. Number of heads and tails are noted. Is this coin biased?Check with 95% Confidence Level.
Heads = 40 & Tails = 60
Ho: Coin is biased. Ha: Coin is not biased.
Alpha = 0.05

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

Flip	Expected	Observed	O-E	$(O-E)^2$	$(O-E)^2/E$
Head	50	40	-10	100	2
Tail	50	60	10	100	2
					$X^2 = 4$

Chi-square calculated = 4
Chi-square critical = 3.84
Conclusion: Chi-square_calc > chi-square_critical then rejects null hypothesis

Example2 :

You work at a nut factory and you're in charge of quality control. The nut factory produces a nut mix that's supposed to be 50% peanuts, 30% cashews, and 20% almonds.

To check that the nut mix proportions are acceptable, you randomly sample 1000 nuts and find the following frequencies:

Nut	**Frequency**
Peanuts	621
Cashew	189
Almonds	190

Should you reject the null hypothesis that the nut mix has the desired proportions of nuts?

a) I should reject the null hypothesis.
b) I should fail to reject the null hypothesis

Sol:
Correct Answer = a - I should reject the null hypothesis.

Step 1: Calculate the expected frequencies

Nut	Observed	Expected
Peanuts	621	1000 * 0.5 = 500
Cashew	189	1000 * 0.3 = 300
Almonds	190	1000 * 0.2 = 200

Step 2: Calculate chi-square

Phenotype	Observed	Expected	$O - E$	$(O - E)^2$	$(O - E)^2$ / E
Peanuts	621	500	121	14 641	29.28
Cashew	189	300	-111	12 321	41.07
Almonds	190	200	-10	100	0.5

$X^2 = 29.28 + 41.07 + 0.5 = 70.85$

Step 3: Find the critical chi-square value

Since there are three groups, there are two degrees of freedom.

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

Step 4: Compare the chi-square value to the critical value

$X^2 = 70.85$

Critical value = 5.99

The X^2 value is greater than the critical value.

Step 5: Decide whether the reject the null hypothesis

The X^2 value is greater than the critical value, so you should **reject** the null hypothesis that the nut mix has the desired proportions of nuts. The data suggests that there's a problem with the nut mix.

Example 3:

After losing a board game, your friend believes she might have lost because of a problem with your dice. To find out, she rolls your dice 60 times and obtains the following frequencies:

Number	Frequency
1	8
2	11
3	6
4	9
5	12
6	14

Should she reject the null hypothesis that the dice lands on each number with equal probability ($p_1 = p_2 = p_3 = p_4 = p_5 = p_6$)?

a) She should reject the null hypothesis.
b) She should fail to reject the null hypothesis.

Solution :

Correct Answer = a - She should fail to reject the null hypothesis.

Step 1: Calculate the expected frequencies

Number	Observed	Expected
1	8	60 * (1/6) = 10
2	11	10
3	6	10
4	9	10
5	12	10
6	14	10

Step 2: Calculate chi-square

Phenotype	Observed	Expected	$O - E$	$(O - E)^2$	$(O - E)^2 / E$
1	8	10	-2	4	0.4

2	11	10	1	1	0.1
3	6	10	-4	16	1.6
4	9	10	-1	1	0.1
5	12	10	2	4	0.4
6	14	10	4	16	1.6

$X^2 = 0.4 + 0.1 + 1.6 + 0.1 + 0.4 + 1.6 = 4.2$

Step 3: Find the critical chi-square value

Since there are six groups, there are five degrees of freedom.

For a test of significance at $\alpha = .05$ and $df = 5$, the X^2 critical value is 11.07.

Step 4: Compare the chi-square value to the critical value

$X^2 = 4.2$
Critical value = 11.07

The X^2 value is less than the critical value.

Step 5: Decide whether the reject the null hypothesis

The X^2 value is greater than the critical value, so your friend should **fail to reject** the null hypothesis that the die lands on each number with equal probability. Based on the data, there's no reason to think there's a problem with the dice.

4.2.10 Annova test

Analysis of variance (ANOVA) is a statistical technique used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.
In simpler and general terms, it can be stated that the ANOVA test is used to identify which process, among all the other processes, is better.

Example: Suppose, there is a group of patients who are suffering from fever. They are being given three different medicines that have the same functionality i.e. to cure fever. To understand the effectiveness of each medicine and choose the best among them, the ANOVA test is used.

Why ANOVA instead of multiple t-tests?
As the number of groups increases, the number of two sample t-tests also increases.
With increases in the number of t-tests, the probability of making **the type 1 error also increases**.
Example- How many T-tests we need to conduct if we have to compare 4 samples.? 6-T-tests

Each test is done with 95% confidence level.6 test result in confidence level of
0.95*0.95*0.95*0.95*0.95*0.95=0.75
So if we conduct 6 t-tests we will make 25% mistakes which is 25 out of 100 samples .
Here, If we use ANOVA with 95% confidence, there are 5% chances of making a mistake

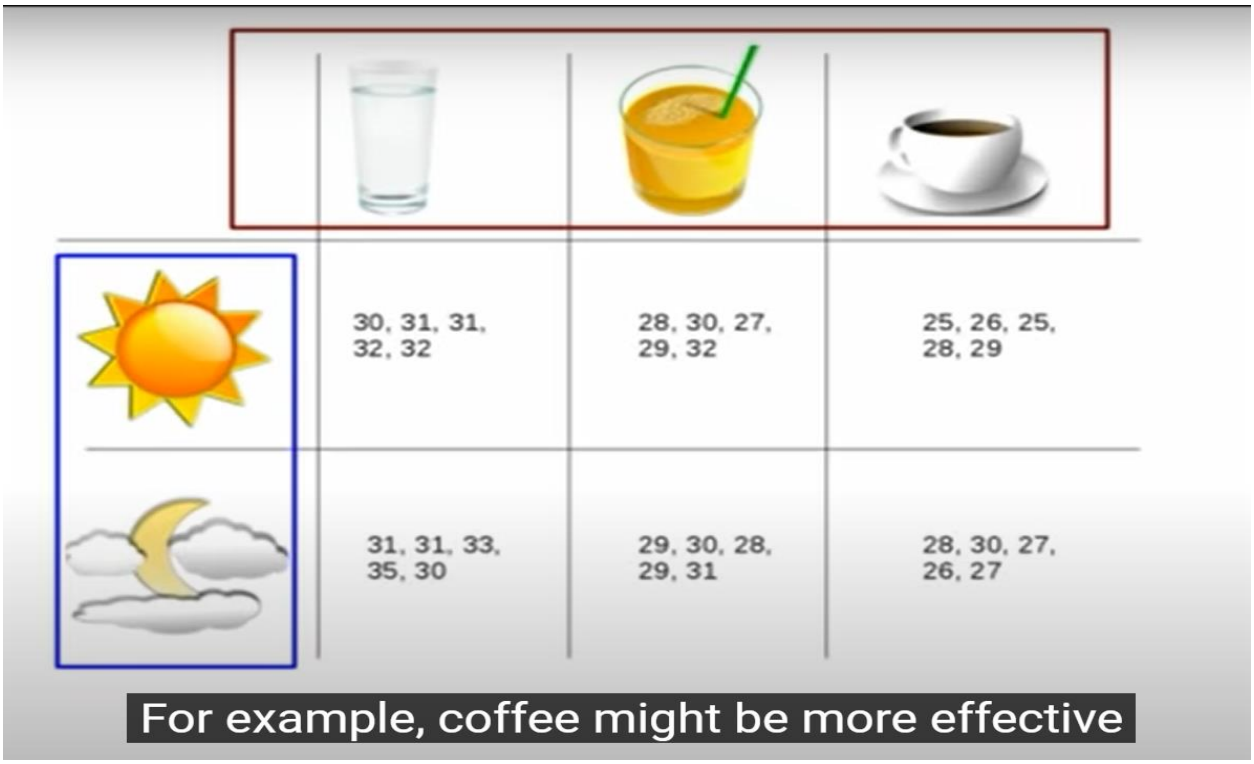
Types of Anova:

1. **One-Way ANOVA:** A test that allows one to make comparisons between the means of three or more groups of data for a single independent categorical variable. With the help of the F-test it helps us to compare the means of three or more samples.
   Example: analyzing the test score of a class based on age(age is binned) . Here test score is a dependent variable and age is the independent variables..

2. **Two-Way ANOVA :** A test that allows one to make comparisons between the means of three or more groups of data, where two independent variables are considered.
   For example: analyzing the test score of a class based on gender and age. Here test score is a dependent variable and gender and age are the independent variables.

3. **N-Way ANOVA (MANOVA) :**

Testing the various beverages(water, juice, cofffe) based on their rection on a set of people in morning and evening



One-factor ANOVA	Two-factors ANOVA
Does a person's place of residence (independent variable) influence his or her salary?	Does the place of residence (independent variable) of a person influence their salary?
Testing the various beverages based on their reaction on a set of people.	Testing the various beverages based on their reaction on a set of people in morning and evening

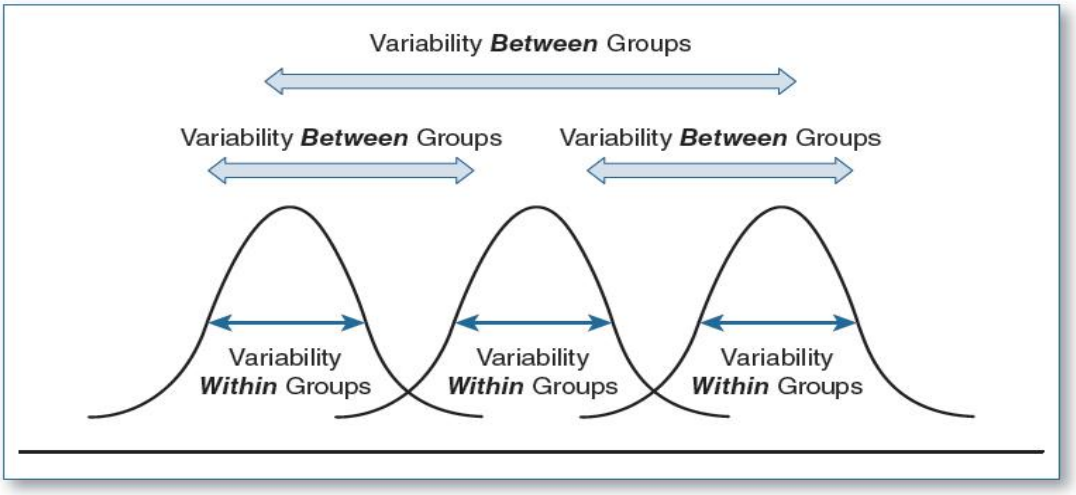
❖ Assumptions of Anova:

The ANOVA test has important assumptions that must be satisfied for the associated p-value to be valid. Source :

- **Independent Sample**: The samples are independent.
- **Normal Distribution:** Each sample is from a normally distributed population.
- The population standard deviations of the groups are all equal. This property is known as homoscedasticity.

- **Terminologies in Anova:** Source

- The **grand mean** is the mean of sample means or the mean of all observations combined, irrespective of the sample.
- The null hypothesis states that all the sample means are equal or the factor did not have any significant effect on the results. Whereas, the alternate hypothesis states that at least one of the sample means is different from another. But we still can't tell which one specifically.
    - H_0= The mean value of all groups is the same and H_1: There are differences in the mean values of the groups



- **Between Group Variability :**

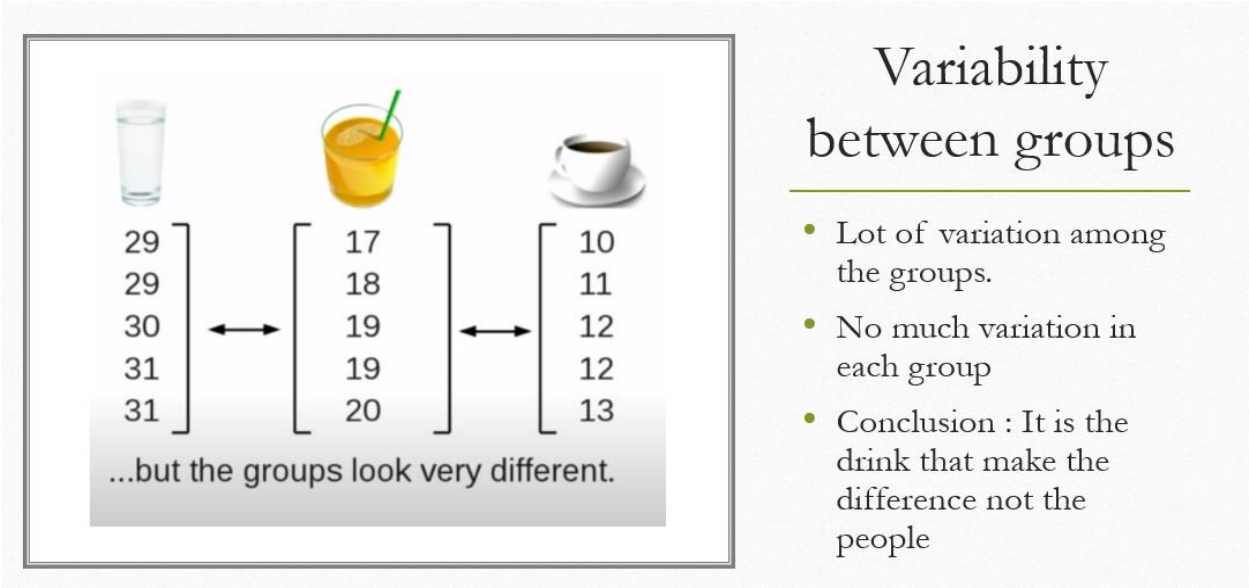
It refers to variations between the distributions of individual groups (or levels) as the values within each group differ.
Each sample is examined, and the difference between its mean and grand mean is calculated to calculate the variability. If the distributions overlap or are close, the grand mean will be similar to the individual means, whereas if the distributions are far apart, the difference between means and grand mean would be large.

We will calculate **Between Group Variability** just as we calculate the standard deviation. First find the sum of each squared deviation and divide it by the degrees of freedom. A deviation is given greater weight if it's from a larger sample. Hence, we'll multiply each squared deviation by each sample size and add them. This is called the **sum-of-squares for between-group variability.**

$$SS_{between} = n_1(\overline{x_1} - \overline{x_G})^2 + n_2(\overline{x_2} - \overline{x_G})^2 + n_3(\overline{x_3} - \overline{x_G})^2 + \ldots. n_k(\overline{x_k} - \overline{x_k})^2$$

For our between-group variability, we will find each squared deviation, weigh them by their sample size, sum them up, and divide by the degrees of freedom (), which in the case of between-group variability is the number of sample means (k) minus 1. Mean square for between -group variability is given by

$$MS_{between} = \frac{n1(\overline{x_1} - \overline{x_G})2 + n2(\overline{x_2} - \overline{x_G})2 + n3(\overline{x_3} - \overline{x_G})2 + ...nk(\overline{x_k} - \overline{x_k})2}{k-1}$$



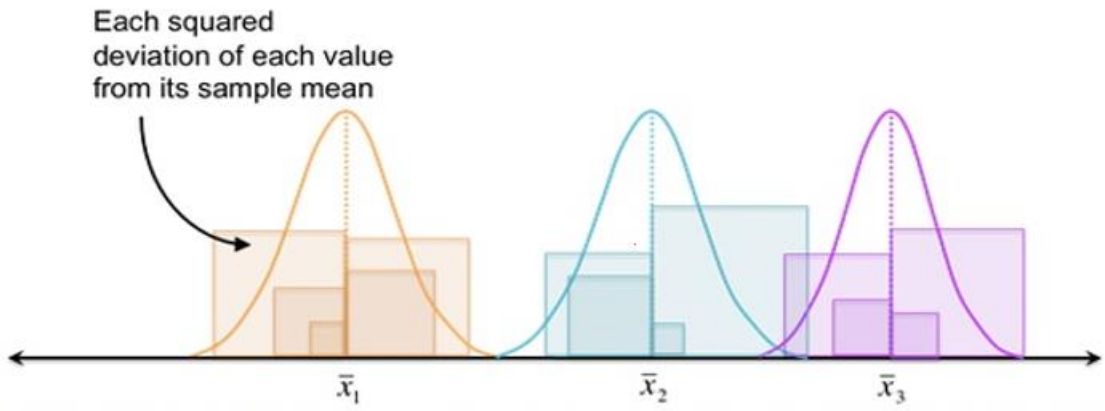
- **Within Group Variability**

  It refers to variations caused by differences within individual groups (or levels), as not all the values within each group are the same.

  We can measure **Within-group variability** by looking at how much each value in each sample differs from its respective sample mean. So first, we'll take the squared deviation of each value from its respective sample mean and add them up. This is the **sum of squares for within-group variability**.

$$SS_{within} = \Sigma(x_{i1} - \bar{x}_1)^2 + \Sigma(x_{i2} - \bar{x}_2)^2 + \ldots + \Sigma(x_{ik} - \bar{x}_k)^2$$
$$= \Sigma(x_{ij} - \bar{x}_j)^2$$

Note: x_{i1} is the ith value from the first sample, x_{i2} is the ith value from the second sample, and so on all the way to x_{ik}, the ith value from the kth sample. x_{ij} is therefore the ith value from the jth sample.



With within-group variability, SS_{within} is the sum of each squared deviation of each value from its respective sample mean (the total area of all the squares in the figure above). MS_{within} is the average-sized square.

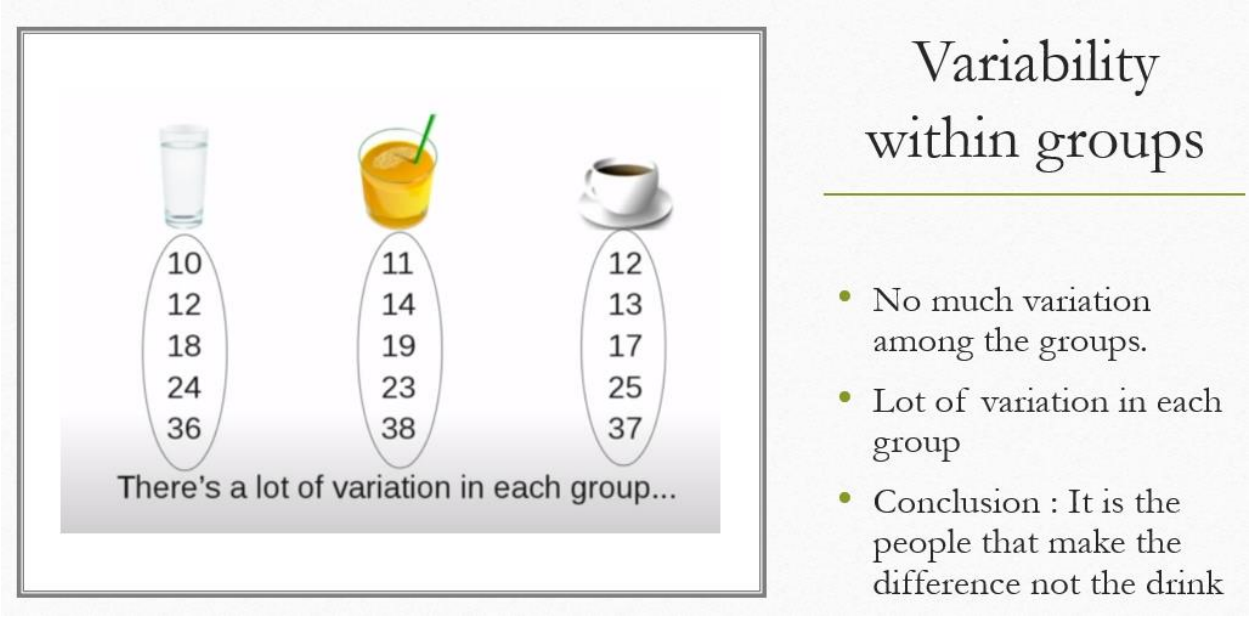
The degrees of freedom is the sum of the sample sizes (N) minus the number of samples (k).

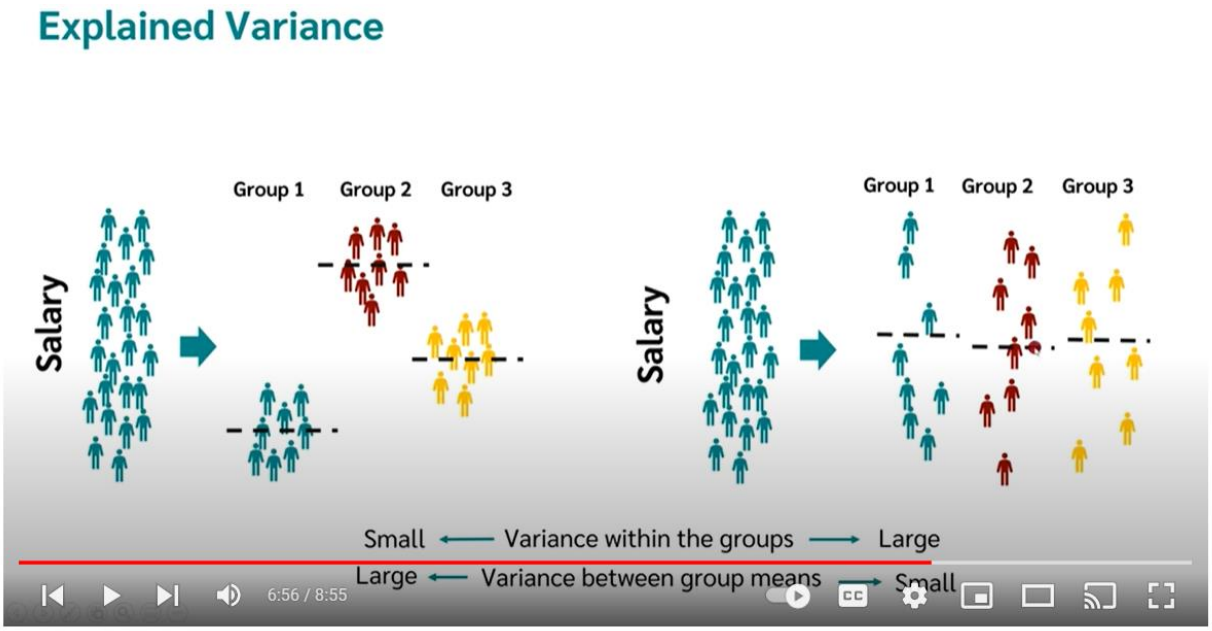
$$df_{within} = (n_1 - 1) + (n_2 - 1) + \ldots (n_k - 1) = n_1 + n_2 + n_3 + \ldots. + n_k - k(1) = N - k$$

The mean square for within-group variability

$$MS_{within} = \sum(x_{ij} - \overline{x_j})^2 / (N - k)$$

Where i = value in each sample,j is sample number , k = number of samples . N is total number of values in all samples.





ANOVA (Analysis of variance) simply explained

- **F-Statistic (F-test)**

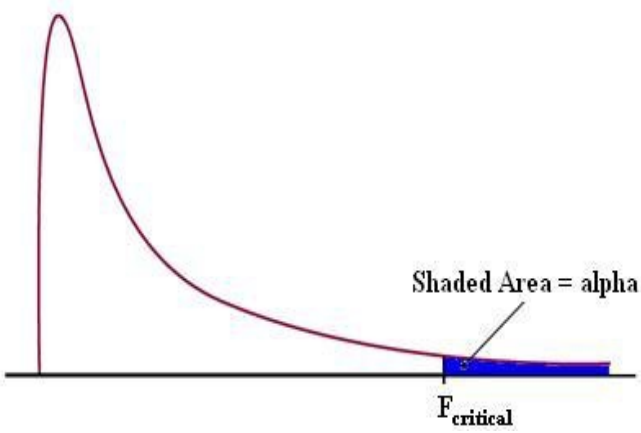
The statistic that measures whether the means of different samples are significantly different is called the F-Ratio. The lower the F-Ratio, the more similar will the sample means be. In that case, we cannot reject the null hypothesis.

$$Fratio = \frac{Mean\ variance\ between\ the\ groups}{Mean\ variance\ within\ the\ groups} = \frac{SS_{btwn}}{SS_{within}} = F(b, w)$$

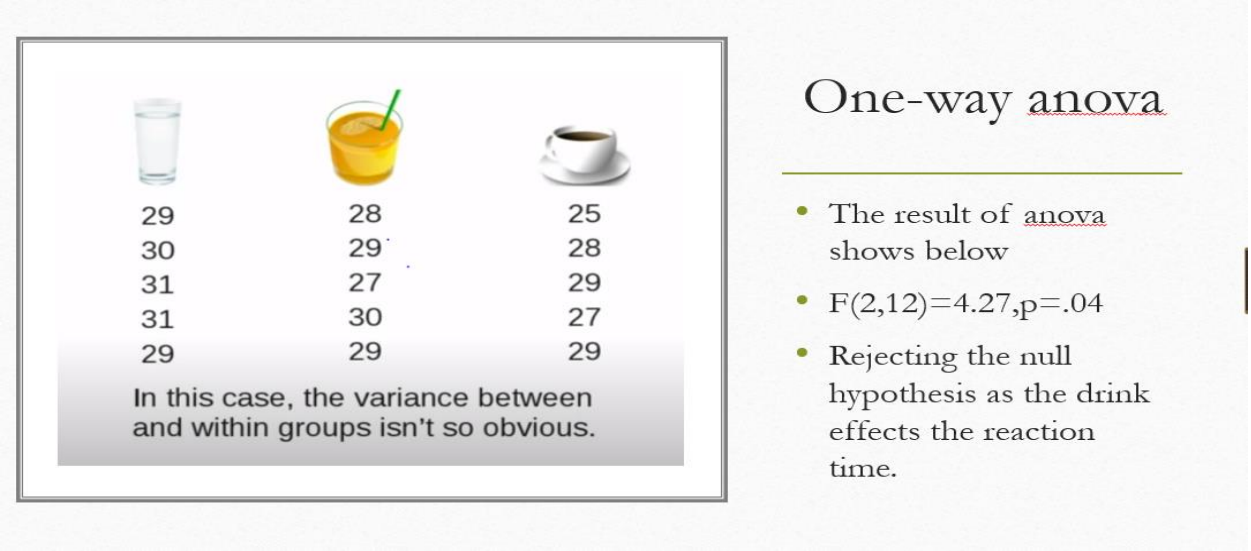
Where b ,w are degree of freedom for between or within groups.

This F-statistic calculated here is compared with the F-critical value for concluding. In terms of our medication example, if the value of the calculated F-statistic is more than the F-critical value (for a specific α/significance level), then we reject the null hypothesis and can say that the treatment had a significant effect. the F-distribution has no negative values because between and within-group variability are always positive due to squaring each deviation.

There is only one critical region in the right tail (shown as the blue-shaded region above). If the F-statistic lands in the critical region, we can conclude that the means are significantly different, and we reject the null hypothesis



F-Table: there are different f-table for various significance values(alpha = 0.01,0.03,0.05)


29 28 25
30 29 28
31 27 29
31 30 27
29 29 29
In this case, the variance between and within groups isn't so obvious.
One-way anova
The result of anova shows below
F(2,12)=4.27,p=.04
Rejecting the null hypothesis as the drink effects the reaction time.

Source

Calculate single-factor analysis of variance

To calculate an analysis of variance, the means of the individual groups and the overall mean must first be calculated. Then the different sums of squares QS can be calculated.

	Group 1	Group 2	Group 3
	3	10	7
	2	9	6
	1	9	7
	1	8	6
	4	7	5
	2	8	4
	4	6	3
	3	7	6
Mean	2.5	8	5.5

Mean value of all data

$$G = \frac{\sum x}{N}$$

$$G = \frac{3 + 10 + \ldots + 7 + 6}{24} = 5.3$$

Mean value of the groups

$$A_i = \frac{\sum^{Group} x}{n_{Group}}$$

$$A_1 = \frac{3 + 2 + \ldots + 4 + 3}{8} = 2.5$$

$$A_2 = \frac{10 + 9 + \ldots + 6 + 7}{8} = 8$$

$$A_3 = \frac{7 + 6 + \ldots + 3 + 6}{8} = 5.5$$

Square sum Total

$$QS_{Total} = \sum (x - G)^2$$

$$QS_{Total} = (3 - 5.3)^2 + (2 - 5.3)^2 + \ldots + (6 - 5.3)^2 = 157.3$$

Square sum between groups

$$QS_{btw} = \sum^{Groups} n_i (A_i - G)^2$$

$$QS_{btw} = 8\left((2.5 - 5.3)^2 + (8 - 5.3)^2 + (5.5 - 5.3)^2\right) = 121.3$$

Square sum within groups

$$QS_{wi} = \sum^{Groups} \sum^{Group} (x_{mi} - A_i)$$

$$QS_{wi} = (3 - 2.5)^2 + (2 - 2.5)^2 + \ldots + (10 - 8)^2 + (9 - 8)^2 + \ldots + (7 - 5.5)^2 + (6 - 5.5)^2 + \ldots = 36$$

Explained Variance

$$\frac{QS_{btw}}{QS_{Total}} = \frac{121.33}{157.33} = 0.77$$

Mean squares

$$MQ_{btw} = \frac{QS_{zw}}{df_{zw}} = \frac{121.33}{2} = 60.67$$

$$MQ_{wi} = \frac{QS_{in}}{df_{in}} = \frac{36}{21} = 1.71$$

F-Value

$$F = \frac{MQ_{zw}}{MQ_{in}} = \frac{60.67}{1.71} = 35.39$$

The mean squares can then be calculated from the square sums and finally the F-value can be calculated. The p-value can then be calculated from the F-value and the degrees of freedom using the F-distribution.