https://lantsandlaminins.com/statistics-test-flow-chart/

Basics of Statistics

There are two types of statistics:

Measure of frequency
Measure of dispersion
Measure of central tendency
Measure of position

1] Descriptive

We use descriptive stats to summarize the data using whatever measurement or data we have and describe the data in stats like mean,median.

2] Inferential

In inferential stats we use sample statistics like mean & stdev to find statistics related to population using different inferential techniques like confidence interval & hypothesis testing.

Descriptive Analytics: tells us what happened in the past and helps a business understand how it is performing by providing context to help stakeholders interpret information.

Predictive Analytics predicts what is most likely to happen in the future and provides companies with actionable insights based on the information.

Measure of central Tendency:

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

Mean: 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.

Median: The middle value of the dataset.it will tell us 50% of the data is above Median value and 50% of the data is below Median Value. **Ex- 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.

Percentiles and Quartiles (Not a measure of variability)

- **Percentiles** — A measure that indicates the value below which a given percentage of observations in a group of observations falls.
- Ex.75% is -->500 : it tell us 75% of the data is below 500
- **Quantiles**— Values that divides the data points of the data into four equal parts(Q1,Q2,Q3,Q4)

Measure of Variability: OR Dispersion

Range: The difference between the highest and lowest value in the dataset.As range doesn't consider each value from the data to calculate range so this is not an ideal measure.

interquartile range is a measure of where the bulk of the values lie

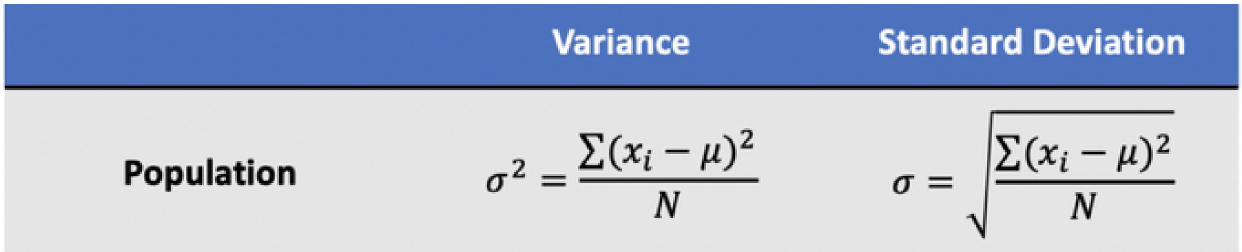
Interquartile Range (IQR)— A measure of statistical dispersion and variability based on dividing a data set into quartiles. IQR = Q3 − Q1 https://towardsdatascience.com/why-1-5-in-iqr-method-of-outlier-detection-5d07fdc82097

Variance: The average squared difference of the values from the mean to measure how spread out a set of data is relative to mean.
It tells us how the data is spread and variation in the data.

Z-Score: The value of the z-score tells you how many standard deviations the value is away from the mean. ... 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-μ)/σ

Standard Deviation: The standard deviation tells us how data is spread out. A low value for the standard deviation indicates that the data points are close to the mean, while a high value for the standard deviation indicates that the data are spread out.



	Variance	Standard Deviation
Population	$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$	$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$

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.

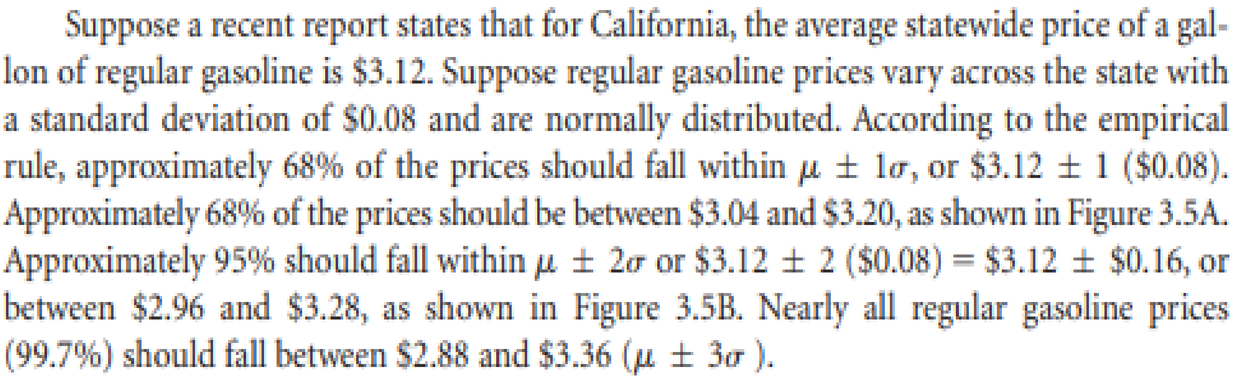
Standard Deviation Example: If we want to see how the volume of perfume bottles is manufactured then in this case if stdev is 2 then we can say 95% bottles volume will be between +- 2 stddev.but variance doesn't make sense here.

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.**stdev is sensitive to outliers.**

Two ways of applying the standard deviation are the empirical rule and Chebyshev’s theorem.

1] 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.

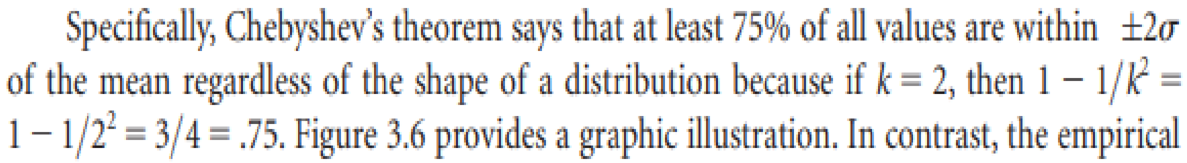


Suppose a recent report states that for California, the average statewide price of a gallon of regular gasoline is \$3.12. Suppose regular gasoline prices vary across the state with a standard deviation of \$0.08 and are normally distributed. According to the empirical rule, approximately 68% of the 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 standard deviations of the mean regardless of the shape of the distribution.**(2stdev -75% & 3stdev - 88%)**



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

Covariance: Covariance is a measure of how much two random variables vary together.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.

Variable measurement scale:
1. Nominal: Categorical or label:
2. ordinal: rank order
3. Interval: Rank order + Equal interval
4. Ratio: Rank order + Equal interval + absolute zero

ratio>interval>ordinal>nominal

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.

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

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

Application of Mean,Mode,Median:

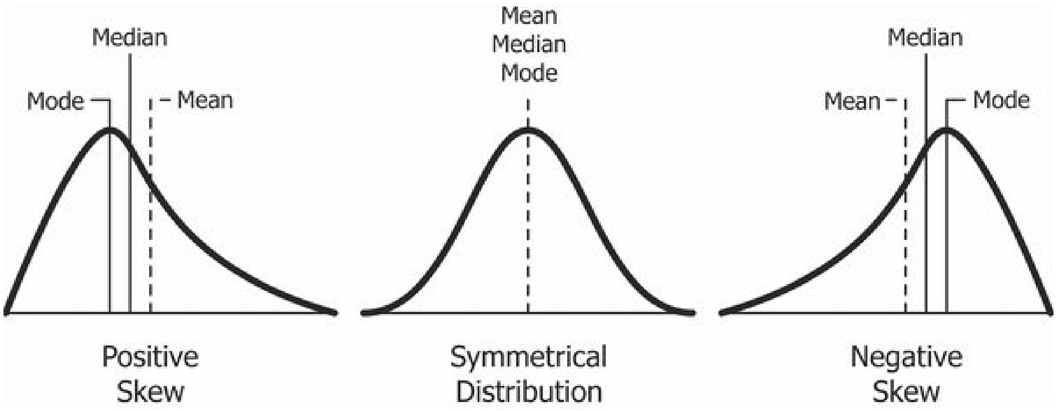
We use them for finding the skewness and kurtosis of the data.

Measure of shape:

1] Skewness:

It tells about the position of the majority of data values in the distribution around the mean value.

Skewness = 3 * (Mean – Median) / Standard Deviation.



1] **Positive Skewness**(Right skew)

Positive Skewness means when the tail on the right side of the distribution is longer or fatter. The mean and median will be greater than the mode.

2] **Negative Skewness**(Left Skew)

Negative Skewness is when the tail of the left side of the distribution is longer or fatter than the tail on the right side. The mean and median will be less than the mode.

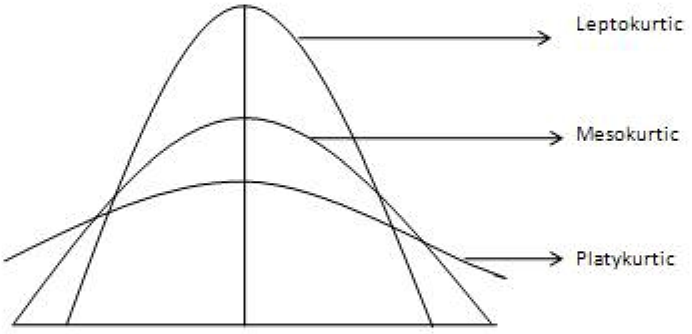
Skewness Example.

Let us take a very common example of house prices. Suppose we have house values ranging from $100k to $1,000,000 with the average being $500,000.

If the peak of the distribution was left of the average value, portraying a *positive skewness* in the distribution. It would mean that many houses were being sold for less than the average value, i.e. $500k. This could be for many reasons, but we are not going to interpret those reasons here.

If the peak of the distributed data was right of the average value, that would mean a *negative skew.* This would mean that the houses were being sold for more than the average value.

2] Kurtosis:



Kurtosis is all about the tails of the distribution — not the peakedness or flatness. It is used to describe the extreme values in one versus the other tail. It is actually the *measure of outliers* present in the distribution.

High kurtosis in a data set is an indicator that data has heavy tails or outliers. If there is a high kurtosis, then, we need to investigate why do we have so many outliers. It indicates a lot of things, maybe wrong data entry or other things. Investigate!

Low kurtosis in a data set is an indicator that data has light tails or lack of outliers. If we get low kurtosis(too good to be true), then also we need to investigate and trim the dataset of unwanted results.

Sampling Methods:

1] Random sampling

Random sampling occurs when every member of a population has an equal chance of being selected to be in the sample

Example: We put the names of every student in a class into a hat and randomly draw out names to get a sample of students.

2] Stratified random sample:

It is based on the assumption that the distribution of the characteristics of interest within each stratum is similar to the overall distribution.

Split a population into groups. Randomly select some members from each group to be in the sample.

Example: Split up all students in a school according to their grade – freshman, sophomores, juniors, and seniors. Ask 50 students from each grade to complete a survey about the school lunches.

3] Cluster random sample:

Split a population into clusters. Randomly select some of the clusters and include all members from those clusters in the sample.

4] Systematic random sample:

Put every member of a population into some order. Choosing a random starting point and select every nth member to be in the sample

EX- A teacher puts students in alphabetical order according to their last name, randomly chooses a starting point, and picks every 5th student to be in the sample.

Central limit theorem:

The central limit theorem says that the sample mean will be approximately normally distributed for larger sample sizes (n), regardless of the distribution from which we are sampling.

According to the central limit theorem, if a population is normally distributed, the sample means for samples taken from that population also are normally distributed regardless of sample size.

From mathematical expectation it can be shown that the **mean of the sample means** is the population mean.

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

The standard deviation of the sample means (called the standard error of the mean) is the standard deviation of the population divided by the square root of the sample size.

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

The real advantage of the central limit theorem comes when sample data drawn from populations which are not normally distributed or from populations of unknown shape also can be analyzed by using the normal distribution(z-test) because the sample means are normally distributed for sufficiently large sample sizes.
The central limit theorem states that sample proportions are normally distributed for large sample sizes.

EX*-If we take 1000 Samples from a population with sample size(N) 30 (30 people in one sample) we will end up with 1000 Mean.if we plot histogram of these 1000 Mean it will be Normally distributed even if the original population is not normally. distributed.as we increase sample size(N) data will be more normally distributed.It is important in hypothesis testing to calculate CI.*

Statistical inference: estimation for single population:

Techniques for estimating population parameters (mean,variance) from sample statistics (mean,variance) are important tools for business. These techniques are required for estimating population means,population proportion and the population variance using Sample statistics(mean,variance).

In business research a product is new or untested or information about the population is unknown. In such cases, gathering data from a sample and making estimates about the population is useful and can be done with a point estimate or an interval estimate.

A sample mean is equal to population mean so this is a point estimate.

A point estimate(mean or variance) is the use of a statistic from the sample as an estimate for a parameter(mean or variance) of the population. Because point estimates vary with each sample, it is usually best to construct an **interval estimate.**

So we will use the Confidence interval as an interval estimate.

An interval estimate is a range of values computed from the sample within which the researcher believes with some confidence that the population parameter lies. Certain levels of confidence seem to be used more than others: 90%, 95%, 98%, and 99%.

1] Confidence Interval

A Confidence Interval is a range of values we are fairly sure our true value lies in with a certain level of confidence.

Confidence interval is used when we need to estimate(calculate) the population parameters using sample statistics.
Hypothesis Testing is used when we just need to test whether sample statistics are matching with population parameters.

Confidence interval and hypothesis testing both are inferential statistics techniques but we use confidence intervals when we need to calculate range of values where actual true value lies in and we use Hypothesis testing to just test whether sample mean is equal to population mean or sample mean is greater than or less than population mean.

a] If the population standard deviation is known, the z statistic is used to estimate the population mean.
b] If the population standard deviation is unknown, the t distribution should be used instead of the z distribution.

Note:

1] Use Z-test when stddev is known and N>30 2] Use T-test when stddev is unknown and N>30 3] Use Z-test when stddev is known & data is normally distributed and N<30 4] Use T-test when stddev is unknown and & data is normally distributed and N<30

$$Best\ Estimate \pm Margin\ of\ Error$$

Where the *Best Estimate* is the **observed population proportion or mean**

$$Population\ Proportion\ or\ Mean\ \pm (t - multiplier * \ Standard\ Error)$$

The Standard Error is calculated differently for population proportion and mean:

$$Standard\ Error\ for\ Population\ Proportion = \sqrt{\frac{Population\ Proportion * (1 - Population\ Proportion)}{Number\ Of\ Observations}}$$

$$Standard\ Error\ for\ Mean = \frac{Standard\ Deviation}{\sqrt{Number\ Of\ Observations}}$$

a] If the population standard deviation is known, the z statistic is used to estimate the population mean.

Z-statistic formula to calculate population mean when the stdev of population is known:

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

Example: calculate the average Height of Pune city

We measure the heights of 40 randomly chosen men, and get a mean height of 175cm.
We also know the standard deviation of men's heights is 20cm using the 95% Confidence Interval.

Here population standard deviation is given so we will use Z-statistic to calculate population average height.

S- standard deviation
n- no of observation
x - sample mean
Z- value for 95% =1.96 , 99%-2.576, 90% --1.645

Based on the following Z-score formula we will calculate the range of population average height we sure actual height will fall..

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

$$175 \pm 1.960 \times \frac{20}{\sqrt{40}}$$

= **175cm ±6.20cm**

In other words: from 168.8cm to 181.2cm

I am 95% confident that all the men's average height will be in the range of 168.8cm to 181.2cm

What does being 95% confident that the population mean is in an interval actually indicate?
It indicates that,, if the researcher were to randomly select 100 samples of 40 people heights and use the results of each sample to construct a 95% confidence interval, approximately 95 of the 100 samples are likely to contain the population mean and 5 samples are not likely to contain the population mean.

b] If the population standard deviation is unknown, the t distribution should be used instead of the z distribution

T-statistic formula to calculate population mean when the stdev of population is unknown:

$$\bar{x} \pm t\frac{s}{\sqrt{n}}$$

In T-distribution t value is calculated based on the degree of freedom and confidence level(alpha).

Degree of freedom = sample size – 1
The degrees of freedom (sometimes abbreviated as d.f.) are the number of free choices left after a sample statistic such as x is calculated.
EX-The number of chairs in a classroom equals the number of students: 25 chairs and 25 students. Each of the first 24 students to enter the classroom has a choice on which chair he or she will sit. There is no freedom of choice, however, for the 25th student who enters the room.

Ex:we have to calculate average overtime work done by employees on weekdays with 90% confidence.
Sample size=18 , df=17 , sample mean=13.56 , sample stdev=7.8
A 90% level of confidence results in = .05 area in each tail.
The table t value is: **t=1.74**

$$13.56 \pm 1.740\frac{7.8}{\sqrt{18}} = 13.56 \pm 3.20$$
$$10.36 \leq \mu \leq 16.76$$

Sample size:For the same sample statistics, as the level of confidence increases, the confidence interval widens.
As we increase sample size the confidence interval decreases.
How large a sample size is needed to guarantee a certain level of confidence for a given margin of error?
By using the formula for the margin of error

$$E = z_c\frac{\sigma}{\sqrt{n}}$$

Given a c - confidence level and a margin of error E, the minimum sample size n needed to estimate the population mean m is

$$n = \left(\frac{z_c\sigma}{E}\right)^2.$$

d] Confidence interval estimation for single population proportion:

The point estimate for p, the population proportion of successes, is given by the proportion of successes in a sample and is denoted by

$$\hat{p} = \frac{x}{n} \qquad \text{Sample proportion}$$

where x is the number of successes in the sample and n is the sample size.

EX-In a survey of 1550 U.S. adults, 1054 said that they use the social media website Facebook. Find a point estimate for the population proportion of U.S. adults who use Facebook.

The number of successes is the number of adults who use Facebook, so x = 1054. The sample size is n = 1550. So, the sample proportion is

$$\begin{aligned} \hat{p} &= \frac{x}{n} && \text{Formula for sample proportion} \\ &= \frac{1054}{1550} && \text{Substitute 1054 for } x \text{ and 1550 for } n. \\ &= 0.68 && \text{Divide.} \\ &= 68\%. && \text{Write as a percent.} \end{aligned}$$

So, the point estimate for the population proportion of U.S. adults who use Facebook is 0.68 or 68%.

How to calculate Sample size for proportion:

$$n = \hat{p}\hat{q}\left(\frac{z_c}{E}\right)^2.$$

You have a preliminary estimate of $\hat{p} = 0.31$. So, $\hat{q} = 0.69$. Using $z_c = 1.96$ and $E = 0.03$, you can solve for n.

$$n = \hat{p}\hat{q}\left(\frac{z_c}{E}\right)^2 = (0.31)(0.69)\left(\frac{1.96}{0.03}\right)^2 \approx 913.02$$

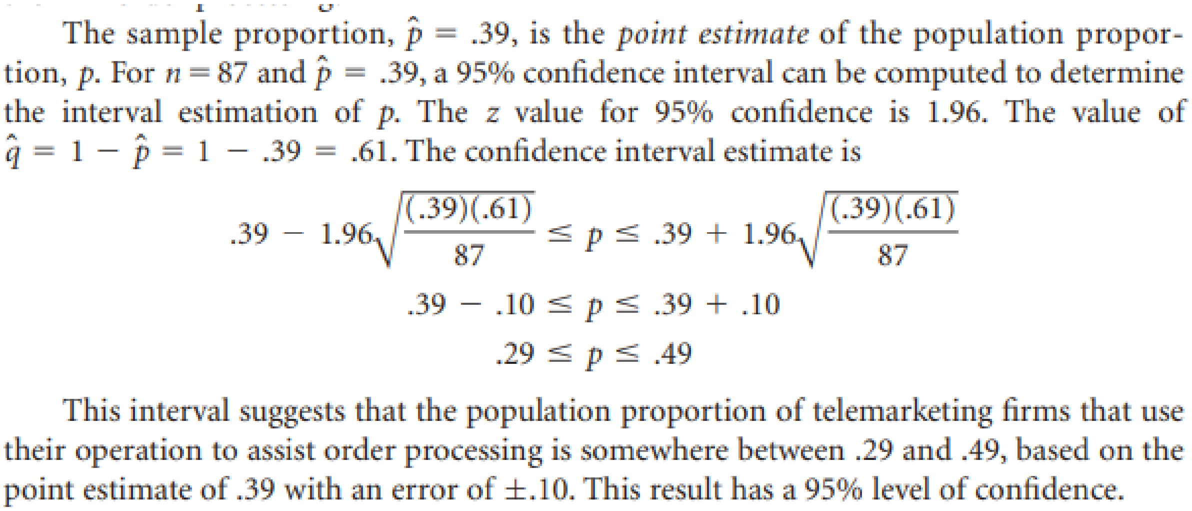
The confidence interval for a population tells us how confident we can be that a sample proportion represents the actual population proportion.

In real life, we usually won’t know the population proportion pp, because we won’t be able to survey or test every subject within our population.

Instead, we’ll have to take a smaller sample of our larger population, and then compute the sample proportion p^. Once we find p^ then
we can use it to make inferences about the value of the population proportion pp.

$$CI = \hat{p} \pm z \times \sqrt{\frac{p(1-p)}{n}}$$

As an example, a study of 87 randomly selected companies with a telemarketing operation revealed that 39% of the sampled companies used telemarketing to assist them in order processing. Using this information, how could a researcher estimate the population proportion of telemarketing companies that use their telemarketing operation to assist them in order processing?



The sample proportion, $\hat{p} = .39$, is the *point estimate* of the population proportion, p. For $n = 87$ and $\hat{p} = .39$, a 95% confidence interval can be computed to determine the interval estimation of p. The z value for 95% confidence is 1.96. The value of $\hat{q} = 1 - \hat{p} = 1 - .39 = .61$. The confidence interval estimate is

$$.39 - 1.96\sqrt{\frac{(.39)(.61)}{87}} \leq p \leq .39 + 1.96\sqrt{\frac{(.39)(.61)}{87}}$$

$$.39 - .10 \leq p \leq .39 + .10$$

$$.29 \leq p \leq .49$$

This interval suggests that the population proportion of telemarketing firms that use their operation to assist order processing is somewhere between .29 and .49, based on the point estimate of .39 with an error of ±.10. This result has a 95% level of confidence.

d] Estimating population variance using Chi square Distribution:

Based on problems sometime we may need to estimate population variance rather than population mean

ex: Parts being used in engines must fit tightly on a consistent basis. A wide variability among parts can result in a part that is too large to fit into its slots or so small that it results in too much tolerance, which causes vibrations. How can variance be estimated?

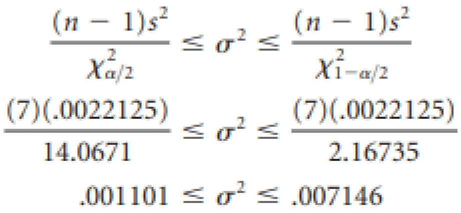
We can use sample variance to estimate population variance because sample variances are typically used as estimates of the population variance.

Chi square(X^2) Formula to calculate variance:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

$$\text{df} = n - 1$$

Chi square(X^2) Confidence interval for estimating population variance:



$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

$$\frac{(7)(.0022125)}{14.0671} \le \sigma^2 \le \frac{(7)(.0022125)}{2.16735}$$

$$.001101 \le \sigma^2 \le .007146$$

df=7, sample variance=0.0022125, chi square value for right tail =14.067 , left tail X^2 2.16

Chi square values are calculated based on degree of freedom and alpha using Chi square table.

Statistical inference: Hypothesis Testing for single population:

What is Hypothesis?
It is an assumption that is made on the basis of some evidence.

What is Hypothesis Testing?
Hypothesis Testing is the process in which we test certain assumptions regarding population. Hypothesis Testing is the process of Rejecting Null Hypothesis or accepting Alternative hypothesis based on the sample data.

All statistical hypotheses consist of two parts, a null hypothesis and an alternative hypothesis. These two parts are constructed to contain all possible outcomes of the experiment or study

In a general approach we try to evaluate two statements(two Hypothesis H0 and H1) on Population(data) using a sample(feature) of data.

Null Hypothesis:

The null hypothesis states that the “null” condition exists; that is, there is nothing new happening, the old theory is still true, the old standard is correct, and the system is in control. H0 is called the Null Hypothesis as it’s our initial assumption.

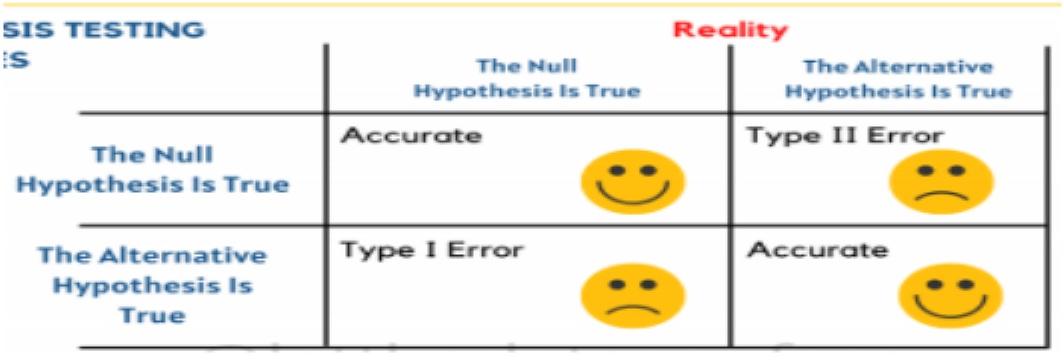
Alternate Hypothesis:

The alternative hypothesis, on the other hand, states that the new theory is true, there are new standards, the system is out of control, and/or something is happening
H1 is called the Alternate Hypothesis.

Type I and Type II error:

A Type I error is committed by rejecting a true null hypothesis.This error is equivalent to False Positive.
A Type II error is committed when a business researcher fails to reject a false null hypothesis. This error is equivalent to False Negative.

https://image2.slideserve.com/3766243/slide20-l.jpg



SIS TESTING :S	Reality The Null Hypothesis Is True	 The Alternative Hypothesis Is True
The Null Hypothesis Is True	Accurate 🙂	Type II Error ☹
The Alternative Hypothesis Is True	Type I Error ☹	Accurate 🙂

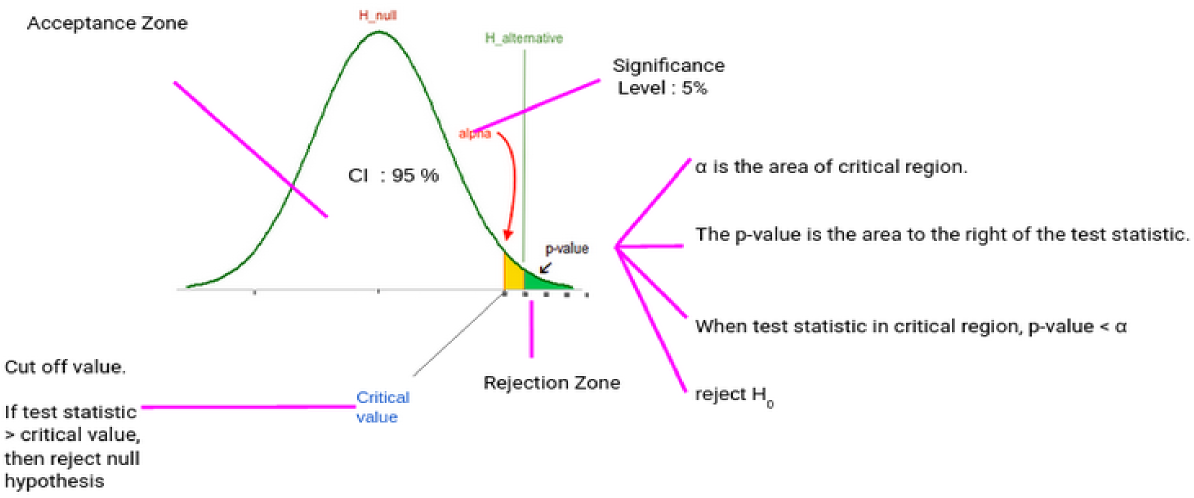
Level of significance(alpha or power):

The probability of committing a Type I error is called alpha or level of significance.

The rejection region is actually dependent on the significance level. The significance level is denoted by α and is the probability of rejecting the null hypothesis if it is true.

The probability of committing a Type II error is beta.

Ideal significance level is 0.05 this 0.05 means that, if we run the experiment 100 times, 5% of the time we will be able to reject the null hypothesis and 95% we will not.



α is the area of critical region.

The p-value is the area to the right of the test statistic.

When test statistic in critical region, p-value < α

reject H_0

If test statistic > critical value, then reject null hypothesis

Critical Value:

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

There are two types of critical values:
1] The critical value can be Z- value(1.96,1.64)
i.e if the calculated z value is greater than 1.96 we can reject null hypothesis.
2] The critical value can be mean value (7k salary)
i.e if calculated salary is greater than 7k we can reject null hypothesis.
p-value has the benefit that we only need one value to make a decision about the hypothesis. We don't need to compute two different values like critical value and test scores.

What is P Value? (https://mlwhiz.com/blog/2019/11/11/pval/)

P-Value is just the probability of observing what we observed in the sample or extreme results if we assume our null hypothesis to be true.
P-value is the observed probability of making type 1 error.we can say p-value is a measure of surprise when p-value is small will have more surprising results and if p_value is large no surprise.

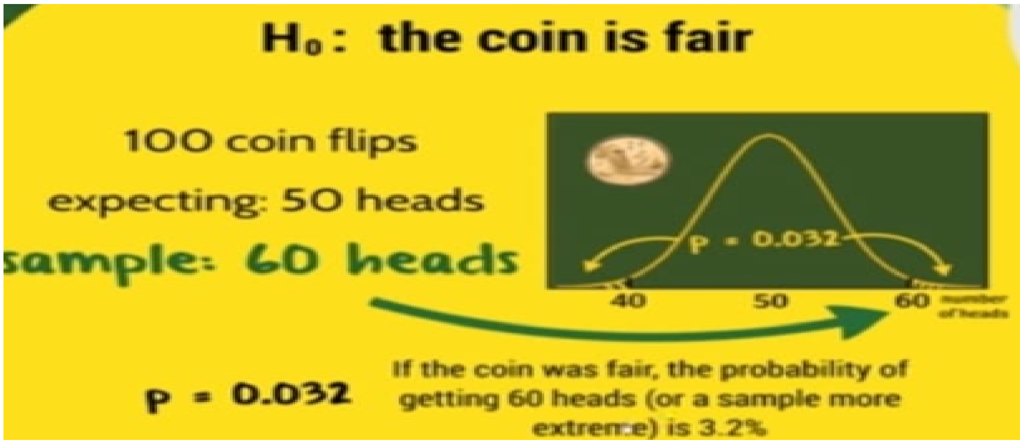
We compare p value with significance level (alpha=0.05)

Assuming the Null hypothesis is true, p-value is the probability of occurring observed results due to chance or randomness. If this probability is small it means observed results are not occured due to chance so there is some probability that this observed value will occur then we can say our null hypothesis is wrong.

In the case of a coin example Assuming null hypothesis is true, if we got 0.01 as p_value then there is only 1% probability of getting 90 heads out of 100 **due to chance** but there is high probability that we can get 90 heads out of 100 from an unfair coin.

If P-value is high it means the result we got are due to chance and If p_value is less it means the result we got are not due to chance.

We can interpret p-value using z- value.let's say for sample statistic you got 2.5 as z value it means sample statistic is 2.5 stdev away from the mean so it indicates there is very less probability of getting that sample statistic if null hypothesis is true.



H₀: the coin is fair

100 coin flips

expecting: 50 heads

sample: 60 heads

p = 0.032

If the coin was fair, the probability of getting 60 heads (or a sample more extreme) is 3.2%

EX-Suppose i have a coin and you are saying that this coin is not fair.so let’s do an hypothesis testing so H0: Coin is fair & Ha: Coin is unfair.
Now if I tossed a single coin 100 times and I got 90 heads then we can say there is something wrong with a coin so we can’t get 90 heads from a fair coin so the coin was not fair.
Now tell me in this example if the coin was fair what is the probability of getting 90 heads or more than that ? So the answer for this question is p_value.

Note: In this case the observed probability is almost 0 which means very less chance that you will reject the true null hypothesis.

We can also say that 90 heads occurred because the coin was unfair and it’s not due to chance.

The test is statistically significant when we reject H0 and imply that there is a specific reason because of which the event has occurred and it is not due to random chance.

The test is statistically insignificant when we fail to reject H0 and imply that the event might have occurred due to random chance alone, and sample data do not provide sufficient evidence for rejecting the null hypothesis.

Procedure for hypothesis Testing:

1] Set up the Null and Alternative Hypothesis. This will enable us to decide whether we have to use a one-tailed (right or left) or a two tailed test.
2] determine the appropriate statistical test and sampling distribution.If the population standard deviation is known and sample size>30 use the z test and if stddev is unknown and sample size is <30 use t-test as the appropriate test statistic.
3] Choose the appropriate Level of Significance ‘α’ depending upon the reliability of the estimates and permissible risk.

4] Collect the data and calculate sample statistics(mean) and based on this data calculate test statistics.
3] Conclusion: We compare the computed value of the test(Z or T) in step 4 with the significant value Z at the given level of significance ‘α’ , and take the decision of either rejecting or fail to reject the null hypothesis. We can also make the decision using the p value. If p-value < α,we can reject the null hypothesis and if p> α fail to reject the null hypothesis.

Example:Testing Hypothesis about population mean using Z statistics

A survey of CPAs across the United States found that the average net income for sole proprietor CPAs is $74,914.* Because this survey is now more than ten years old, an accounting researcher wants to test this figure by taking a random sample of 112 sole proprietor accountants in the United States to determine whether the net income figure changed. The researcher could use the eight steps of hypothesis testing to do so. Assume the population standard deviation of net incomes for sole proprietor CPAs is $14,530.

At step 1, the hypotheses must be established. Because the researcher is testing to determine whether the figure has changed, the alternative hypothesis is that the mean net income is not $74,914. The null hypothesis is that the mean still equals $74,914. These hypotheses follow.

$$H_0: \mu = \$74{,}914$$
$$H_a: \mu \neq \$74{,}914$$

Step 2 is to determine the appropriate statistical test and sampling distribution. Because the population standard deviation is known ($14,530) and the researcher is using the sample mean as the statistic, the z test in formula (9.1) is the appropriate test statistic.

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

Step 3 is to specify the Type I error rate, or alpha, which is .05 in this problem.

Step 4 is to state the decision rule. Because the test is two tails and alpha is .05, there is 2 or .025 area in each of the tails of the distribution. Thus, the rejection region is in the two ends of the distribution with 2.5% of the area in each. There is a .4750 area between the mean and each of the critical values that separate the tails of the distribution (the rejection region) from the nonrejection region. By using this .4750 area and Table A.5, the critical z value can be obtained.

$$z_{\alpha/2} = \pm 1.96$$

displays the problem with the rejection regions and the critical values of z. The decision rule is that if the data gathered produce a z value greater than 1.96 or less than -1.96, the test statistic is in one of the rejection regions and the decision is to reject the null hypothesis. If the observed z value calculated from the data is between -1.96 and +1.96, the decision is to not reject the null hypothesis because the observed z value is in the nonrejection region.

Step 5 is to gather the data. Suppose the 112 CPAs who respond produce a sample mean of \$78,695. At step 6, the value of the test statistic is calculated by using n n= 112, stddev= \$14,530, and a hypothesized mean= \$74,914:

$$z = \frac{78{,}695 - 74{,}914}{\frac{14{,}530}{\sqrt{112}}} = 2.75$$

ACTION: Because this test statistic, z = 2.75, is greater than the critical value of z in the upper tail of the distribution, z = +1.96, the statistical conclusion reached at step 7 of the hypothesis-testing process is to reject the null hypothesis. The calculated test statistic is often referred to as the observed value. Thus, the observed value of z for this problem is 2.75 and the critical value of z for this problem is 1.96.

Note: If Null hypothesis is True it’s very rare to get \$78,696 as avg salary so it means avg salary have changed.

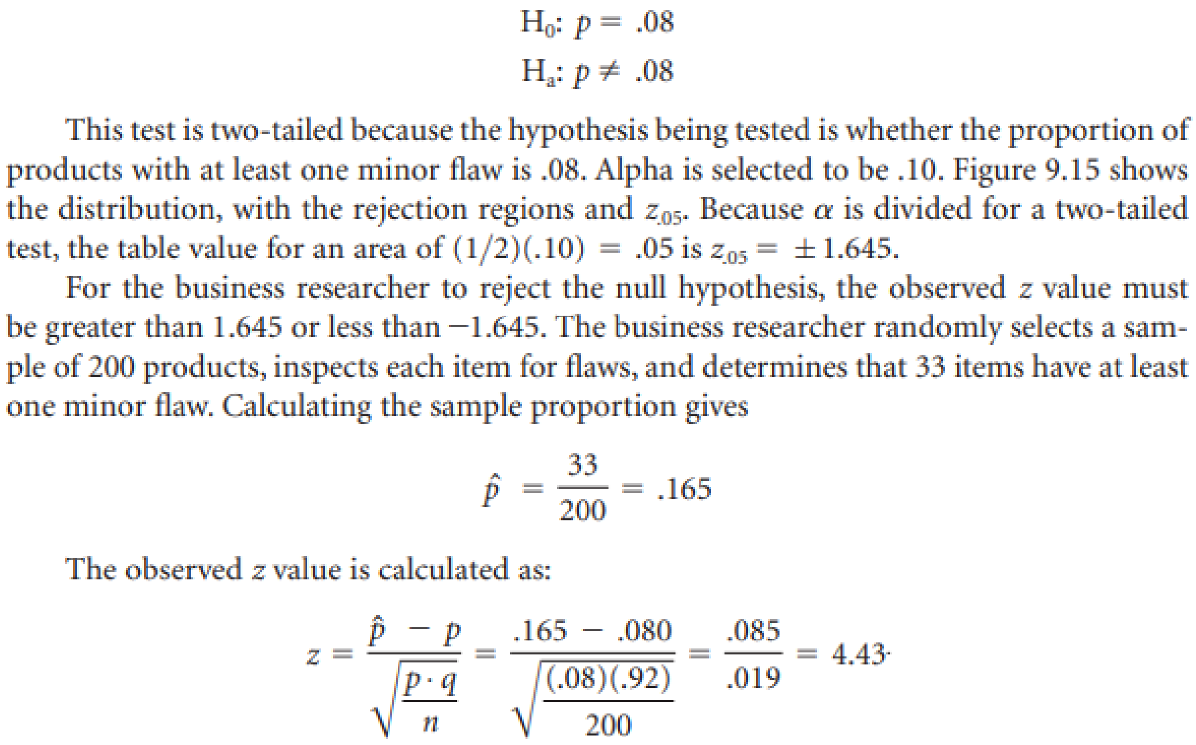
2] **Testing Hypothesis about one population proportion using Z statistics**

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

where

$\hat{p} =$ sample proportion
$p =$ population proportion
$q = 1 - p$

EX- A manufacturer believes exactly 8% of its products contain at least one minor flaw. Suppose a company researcher wants to test this belief. The null and alternative hypotheses are.



$$H_0\text{: } p = .08$$
$$H_a\text{: } p \neq .08$$

This test is two-tailed because the hypothesis being tested is whether the proportion of products with at least one minor flaw is .08. Alpha is selected to be .10. Figure 9.15 shows the distribution, with the rejection regions and $z_{.05}$. Because α is divided for a two-tailed test, the table value for an area of $(1/2)(.10) = .05$ is $z_{.05} = \pm 1.645$.

For the business researcher to reject the null hypothesis, the observed z value must be greater than 1.645 or less than −1.645. The business researcher randomly selects a sample of 200 products, inspects each item for flaws, and determines that 33 items have at least one minor flaw. Calculating the sample proportion gives

$$\hat{p} = \frac{33}{200} = .165$$

The observed z value is calculated as:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.165 - .080}{\sqrt{\frac{(.08)(.92)}{200}}} = \frac{.085}{.019} = 4.43$$

The calculated test statistic 4.43>1.645 so we can reject the null hypothesis.

Statistical inference: Hypothesis Testing & Confidence interval for two population(two independent samples):

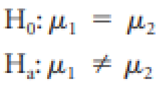
Hypothesis testing and confidence interval about differences in two means using Z-statistic (population variance is known):

1] Hypothesis Testing:

A] Difference in Population Means

In many instances, a business researcher wants to test the differences in the mean values of two populations

As a specific example, suppose we want to conduct a hypothesis test to determine whether the average annual salary for an advertising manager is different from the average annual salary of an auditing manager. Because we are testing to determine whether the means are different, it might seem logical that the null and alternative hypotheses would be



$$H_0\text{: } \mu_1 = \mu_2$$
$$H_a\text{: } \mu_1 \neq \mu_2$$

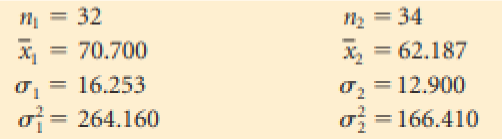
Note, however, that a business researcher could be interested in testing to determine if there is, difference in their salary then we can rewrite our hypothesis like:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

It means if the salary is same for both of them then the difference between salary means is zero.
If the difference between salary means is not zero then salary might be greater than or less than.so that is why this is two tailed test.
The Z-Test formula:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Random samples of 32 advertising managers from across the United States is taken.
A similar random sample is taken of 34 auditing managers.
Mean of the both sample and standard deviation of the two population is given as fallow:



$n_1 = 32$	$n_2 = 34$
$\bar{x}_1 = 70.700$	$\bar{x}_2 = 62.187$
$\sigma_1 = 16.253$	$\sigma_2 = 12.900$
$\sigma_1^2 = 264.160$	$\sigma_2^2 = 166.410$

Suppose alpha= .05. Because this test is a two-tailed test, each of the two rejection regions has an area of .025, leaving .475 of the area in the distribution between each critical value and the mean of the distribution.
The Z-test value is 2.35:

$$z = \frac{(70.700 - 62.187) - (0)}{\sqrt{\frac{264.160}{32} + \frac{166.410}{34}}} = 2.35$$

The observed value of 2.35 is greater than the critical value obtained from the z table, 1.96. The business researcher rejects the null hypothesis and can say that there is a significant difference between the average annual salary of an advertising manager and the average annual salary of an auditing manager.
We can conclude that advertising managers earn more, on the average, than do auditing managers.

B] Hypothesis testing about differences in two proportion:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

EX-Is there a significant difference between the population proportions of parents of black children and parents of Hispanic children who report that their child has had some swimming lessons?

Populations: All parents of black children age 6-18 and all parents of Hispanic children age 6-18

Parameter of Interest: p1 - p2, where p1 = black and p2 = hispanic

Null Hypothesis: p1 - p2 = 0

Alternative Hypothesis: p1 - p2 ≠0

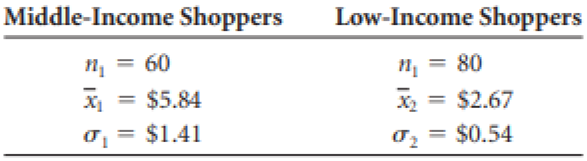
2] Confidence Intervals:

A] confidence interval estimation for Difference of two population

Sometimes being able to estimate the difference in the means of two populations is valuable. By how much do two populations differ in size or weight or age? By how much do two products differ in effectiveness? Do two different manufacturing or training methods produce different mean results? The answers to these questions are often difficult to obtain through census techniques. The alternative is to take a random sample from each of the two populations and study the difference in the sample means.
Confidence Interval Formula for calculating difference between mean:

$$(\bar{x}_1 - \bar{x}_2) - z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + z\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

EX: Suppose a study is conducted to estimate the difference between middle-income shoppers and low-income shoppers in terms of the average amount saved on grocery bills per week by using coupons.
Random samples of 60 middle-income shoppers and 80 low income shoppers are taken.



Middle-Income Shoppers	Low-Income Shoppers
$n_1 = 60$	$n_1 = 80$
$\bar{x}_1 = \$5.84$	$\bar{x}_2 = \$2.67$
$\sigma_1 = \$1.41$	$\sigma_2 = \$0.54$

This information can be used to construct a 98% confidence interval to estimate the difference between the mean amount saved with coupons by middle-income shoppers and the mean amount saved with coupons by low-income shoppers.

$$(5.84 - 2.67) - 2.33\sqrt{\frac{1.41^2}{60} + \frac{0.54^2}{80}} \le \mu_1 - \mu_2 \le (5.84 - 2.67) + 2.33\sqrt{\frac{1.41^2}{60} + \frac{0.54^2}{80}}$$
$$3.17 - 0.45 \le \mu_1 - \mu_2 \le 3.17 + 0.45$$
$$2.72 \le \mu_1 - \mu_2 \le 3.62$$

There is a 98% level of confidence that the actual difference in the population mean coupon savings per week between middle-income and low-income shoppers is between $2.72 and $3.62. That is, the difference could be as little as $2.72 or as great as $3.62.
The point estimate for the difference in mean savings is $3.17. Note that a zero difference in the population means of these two groups is unlikely, because zero is not in the 98% range.

3] We can estimate the confidence interval for comparing two independent proportions.

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}$$

Ex-we can estimate the difference between female and male smoking rate

Note:Use T-test when population variance is unknown for **Hypothesis testing and confidence interval about differences in two means .(population variance is unknown):**

Statistical inference for two related population(two dependent samples):

In the preceding section, hypotheses were tested and confidence intervals constructed about the difference in two population means when the samples are independent. In this section, a method is presented to analyze dependent samples or related samples.
Some researchers refer to this test as the matched-pairs test. Others call it the t-test for related measures or the correlated t test.

What are some types of situations in which the two samples being studied are related or dependent?
Let's begin with the before-and-after study. Sometimes as an experimental control mechanism, the same person or object is measured both before and after a treatment. Certainly, the after measurement is not independent of the before measurement because the measurements are taken on the same person or object in both cases

Basics of Probability:

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

Trial and Event: The performance of an experiment is called a trial, and the set of its outcomes is termed an event.

Ex: Tossing a coin and getting head is a trial. Then the event is {HT, TH, HH}

Random Experiment: It is an experiment in which all the possible outcomes of the experiment are known in advance. But the exact outcomes of any specific performance are not known in advance.

Ex:

1. Tossing a Coin
2. Rolling a die
3. Drawing a card from a pack of 52 cards.
4. Drawing a ball from a bag.

Outcome: The result of a random experiment is called an Outcome.

Ex: 1. Tossing a coin is an experiment and getting head is called an outcome.
2. Rolling a die and getting 6 is an outcome.

Sample Space: The set of all possible outcomes of an experiment is called sample space and is denoted by S.

Impossible Events: An event which will never happen.

Ex: Tossing double-headed coins and getting tails in an impossible event.

Sure Outcome/Certain Outcome: An Outcome which will definitely be happen

Ex: Tossing double-headed coins and getting heads only.

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.

Ex: 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 or Disjoint Events: Events are called mutually exclusive if they cannot occur simultaneously.

Ex: Suppose a card is drawn from a pack of cards, then the events of getting a jack and getting a king are mutually exclusive because they cannot occur simultaneously.

Exhaustive Events: The total number of all possible outcomes of an experiment is called exhaustive events.

Ex: In the tossing of a coin, either head or tail may turn up. Therefore, there are two possible outcomes. Hence, there are two exhaustive events in tossing a coin.

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.

Types of Probability

1] Classical probability:Classical (or theoretical) probability is used when each outcome in a sample space is equally likely to occur. The classical probability for an event E is given by:

P(E) = Number of outcomes in event E / Total number of outcomes in sample space
EX: 1] rolling a die
2]Tossing a coin

2] Empirical (or statistical) probability: is based on observations obtained from probability experiments. The empirical probability of an event E is the relative frequency of event E.

P(E) = Frequency of event E / Total frequency
= f / n
Note that n = sum(f)

Law of large numbers:

As an experiment is repeated over and over, the empirical probability of an event approaches the theoretical (actual) probability of the event.

EX-Suppose if you toss a coin 10 times and you get 3/10 heads Because you tossed the coin only a few times, your empirical probability is not representative of the theoretical probability, which is 50%. So if you repeat this trial 150 times you will almost get 75/100 heads which is an actual(theoretical) probability of 50% .

Subjective probability: Subjective probabilities result from intuition, experience,educated guesses, and estimates. For instance, given a patient’s health and extent of injuries, a doctor may feel that the patient has a 90% chance of a full recovery

Joint probability: Joint probability is a type of probability where more than one event can occur simultaneously. The joint probability is the probability that event A will occur at the same time as event B.

Marginal probability: A probability of any single event occurring unconditioned on any other events.Whenever someone asks you whether the weather is going to be rainy or sunny today(without any conditional or prior information), you are computing a marginal probability.

Conditional probability: is a probability of an event given that (by assumption, presumption, assertion, or evidence) another event has occurred.

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)   . Events A and B are dependent.
2] The probability that two independent events A and B will occur in sequence is
P(A and B) = P(A) * P(B)   . Events A and B are independent.

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

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:

$$P(4 \text{ or ace}) = P(4) + P(\text{ace}) = \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13} \approx 0.154.$$

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

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

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

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

Random Variable:

Random variable is variable whose possible values are numerical outcomes of random experiment.

Random variable is a set of possible values from random experiment.
Random variable is a function that maps events(from sample space S) to the real numbers.

$$X: S \longrightarrow R$$

The word random indicates that x is determined by chance

The outcome of a probability experiment is often a count or a measure. When this occurs, the outcome is called a random variable.
A random variable x represents a value associated with each outcome of a probability experiment.

Types of Random Variables:

1] Discrete random variable
2] Continuous Random variable
3] Mixed random variable

1] **Discrete random variable**

A rv is discrete if its set of possible values are countable.
Ex.1] rolling a dice: Let X be the random variable getting all possible values [1,2,3,4,5,6]
2] No of cars in parking lot
3] no of student failed the exam

In most applications, discrete random variables represent counted data, while continuous random variables represent measured data

2] **Continuous random variable**

A random variable is continuous when it has an uncountable number of possible outcomes, represented by an interval on a number line.
Ex. Let Y be a random variable that is equal to the height of different people [172.99,172.992]
2] weight,

Discrete probability Distribution:

Probability distribution specifies the probabilities for each value that the random variable may take.

Probability distribution of discrete random variable:

If you use a probability function to describe any discrete probability distribution, you will consider the function as a **probability mass function(PMF)**.
A PMF is a function that will always return the probability of the outcome. Therefore, this function is written as: f(x) = P(X=x)

Function f, depicting the probability mass function, will return the probability that a random variable will take the value “x.” So, let us go back to our previous example where we roll a six-sided dice.

In this case, the probability mass function will only return the probability of obtaining a number. For example, if you want to calculate the probability of a dice rolling a 6, you will write the probability mass function as f(6) = 1/6.

PMF(probability mass function):

$$p(x) = P(X = x)$$

The probability of x = the probability(X = one specific x)

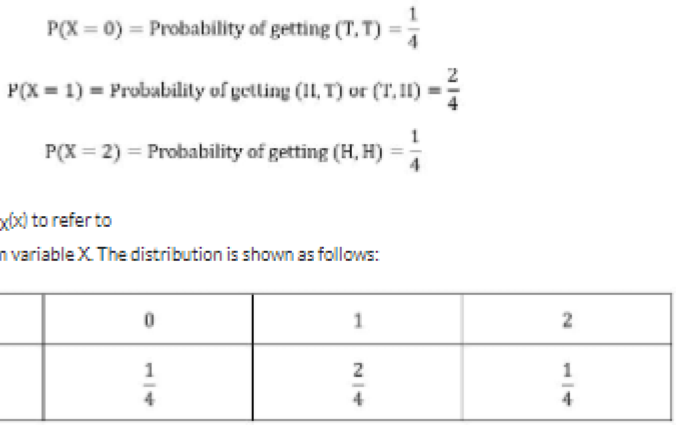
Properties of probability mass function:

1) PMF can never be more than 1 or negative i.e.,

$$0 \leq P_X(x) \leq 1$$

2) PMF must sum to one over the entire range set of a random variable.

$$\sum_{-\infty}^{\infty} P_X(x) = \sum_{x \in R_x} P_X(x) = 1$$



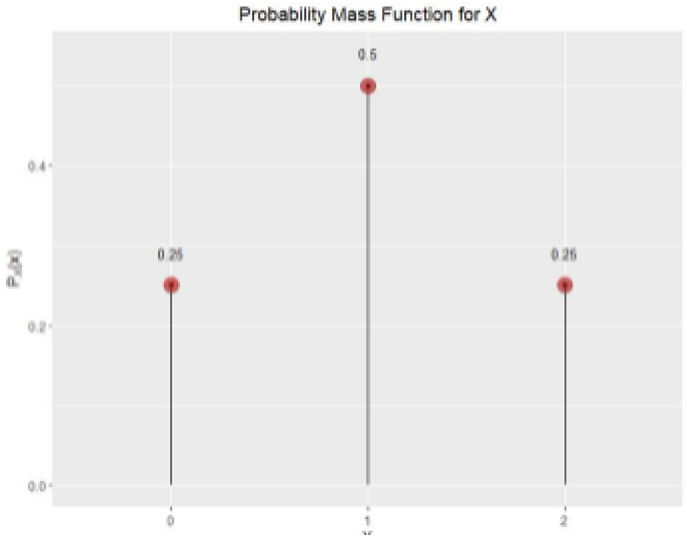
$$P(X = 0) = \text{Probability of getting } (T, T) = \frac{1}{4}$$

$$P(X = 1) = \text{Probability of getting } (H, T) \text{ or } (T, H) = \frac{2}{4}$$

$$P(X = 2) = \text{Probability of getting } (H, H) = \frac{1}{4}$$

We use the notation $P_X(x)$ to refer to the PMF of the random variable X. The distribution is shown as follows:

X	0	1	2
$P_X(x)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$



Mean,Variance and Standard Deviation of Discrete random Variable:

1] Mean

The mean of a discrete random variable is given by

$$\mu = \Sigma x P(x).$$

Each value of x is multiplied by its corresponding probability and the products are added

The mean of a random variable represents the "theoretical average" of a probability experiment and sometimes is not a possible outcome. If the experiment were performed many thousands of times, then the mean of all the outcomes would be close to the mean of the random variable.

Variance and Standard Deviation:

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

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

The **standard deviation** is

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

Expectation or Expected Value:

The mean of a random variable represents what you would expect to happen over thousands of trials. It is also called the expected value.

The expected value of a discrete random variable is equal to the mean of the random variable.

$$\text{Expected Value} = E(x) = \mu = \Sigma x P(x)$$

There are three Discrete probability Distributions:

1] Binomial Distribution
2] Poisson Distribution
3] Geometric distribution

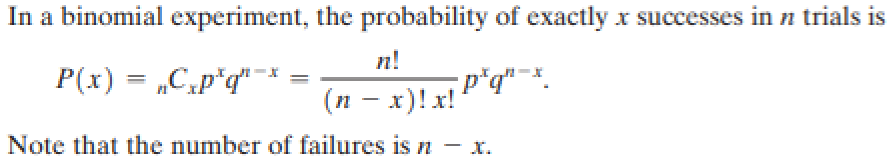
1] Binomial Distribution:

The most widely used discrete distribution is binomial distribution.
As the word binomial indicates, any single trial of a binomial experiment contains only two possible outcomes. These two outcomes are labeled success or failure.

A binomial experiment is a probability experiment that satisfies these conditions:
1] The experiment has a fixed number of trials, where each trial is independent of the other trials.
2] 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).
3] The probability of success is the same for each trial.
4] 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:

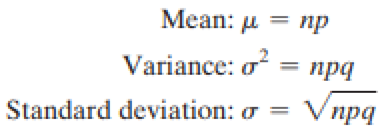


In a binomial experiment, the probability of exactly x successes in n trials is

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

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

Mean , Variance and Standard Deviation of binomial distribution:



$$\text{Mean: } \mu = np$$
$$\text{Variance: } \sigma^2 = npq$$
$$\text{Standard deviation: } \sigma = \sqrt{npq}$$

EX- 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.

2] Poisson Distribution:

.

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.

1. 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.
2. The probability of the event occurring is the same for each interval.
3. 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 m is the mean number of occurrences per interval unit.

EX: Number of telephone calls per minute at a small business

Note: If the number of arrivals per interval is too frequent, the time interval can be reduced enough so that a rare number of occurrences is expected. Another example of a Poisson distribution is the number of random customer arrivals per five-minute interval at a small boutique on weekday mornings.

If a Poisson-distributed phenomenon is studied over a long period of time, a long-run average can be determined. This average is denoted meu( )

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

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

3] 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.

1] A trial is repeated until a success occurs.
2] The repeated trials are independent of each other.
3] The probability of success p is the same for each trial.
4] 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] 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.

Continuous random variable distributions:

a continuous random variable has an infinite number of possible values that can be represented by an interval on a number line. Its probability distribution is called a continuous probability distribution

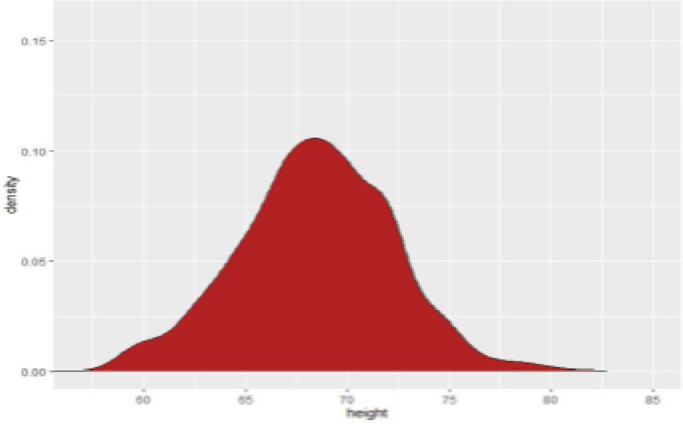
Probability distribution of Continuous random variable:

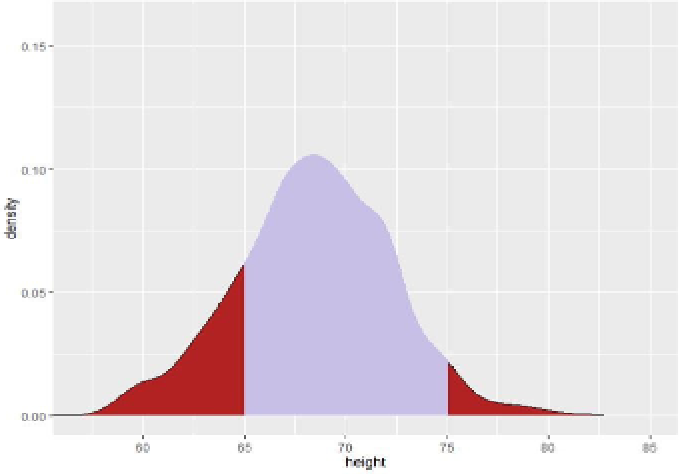
Probability distribution of a continuous random variable is called a probability density function**(PDF)** and it gives probabilities on y-axis.

PDF(probability density function):

$$F(x) = P(a \leq x \leq b) = \int_a^b f(x)dx \geq 0$$

Ex: The distribution of the heights can be shown by a density histogram as follows:

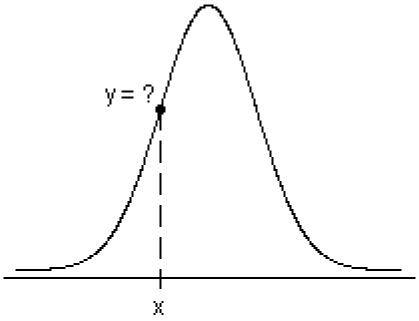




Graphically, the probability that a continuous random variable X takes a value within a given interval is the area below the PDF for X, enclosed between the given interval.

For instance, in the above example, if we wish to determine the probability that a randomly selected person from the population has a height between 65 cm and 75 cm, we calculate the purple area (using definite integration):

Note: In PDF we can’t take exact probability of the height because probability of the exact height will be the line not area under the interval. That's why we have to take intervals of probability for height.



Cumulative Distribution function of Random Variables(CDF):

CDF is used to describe both discrete as well as Continuous random variable distributions.
A CDF shows the probability that a random variable X takes a value lesser than or equal to x.
Mathematically, a CDF is defined as follows:

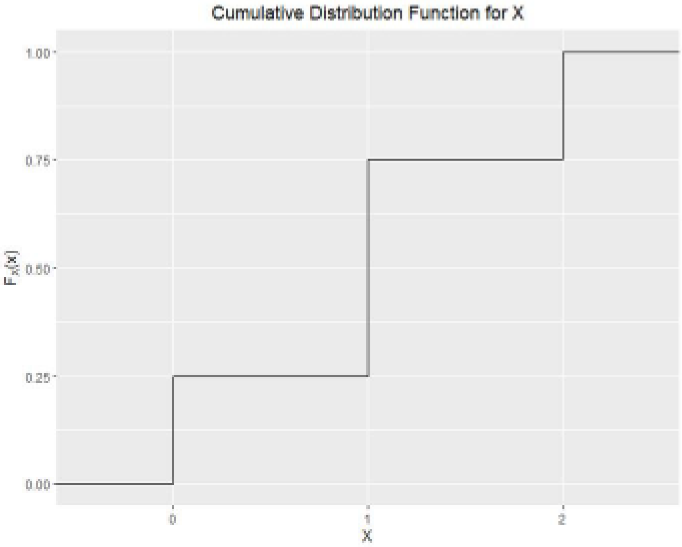
$$F_X(x) = P(X \leq x)$$

For example1, if X is the height of a person selected at random then F(x) is the chance that the person will be shorter than x. If F(180 cm)=0.8. then there is an 80% chance that a person selected at random will be shorter than 180 cm (equivalently, a 20% chance that they will be taller than 180cm)

CDF for Discrete random variable:
ex2: flipping the coin like the above example in PMF:

$$F_X(x) = \left\{ \begin{array}{ll} 0, & \text{for } -\infty < x < 0 \\ \frac{1}{4}, & \text{for } 0 \leq x < 1 \\ \frac{3}{4}, & \text{for } 1 \leq x < 2 \\ 1, & \text{for } 2 \leq x < \infty \end{array} \right\}$$

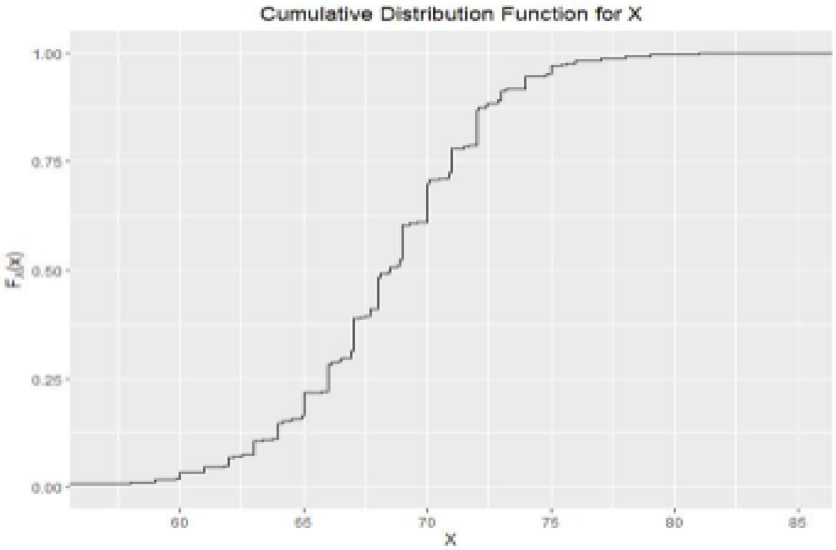
The CDF can also be shown graphically as follows:



CDF(cumulative density function) for Continuous random variable:

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:



Continuous Probability Distributions:

B] 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.

A normal distribution has these 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:

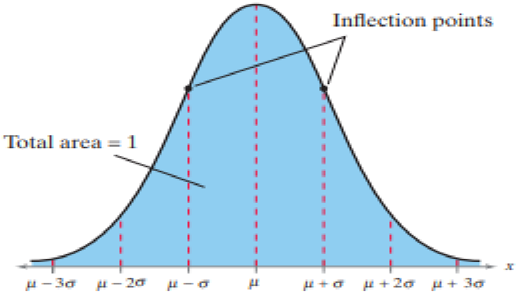
There are infinitely many normal distributions, each with its own mean and standard deviation. The normal distribution with a mean of 0 and a standard deviation of 1 is called the **standard normal distribution**. The horizontal scale of the graph of the standard normal distribution corresponds to z-scores.

z-score is a measure of position that indicates the number of standard deviations a value lies from the mean.

$$\begin{aligned} z &= \frac{\text{Value} - \text{Mean}}{\text{Standard deviation}} \\ &= \frac{x - \mu}{\sigma}. \end{aligned}$$

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 other hand, a standard normal distribution has always the fixed mean and standard deviation.

a] Uniform distribution:

The simplest probability distribution is the uniform distribution, which gives the same probability to any points of a set.

The uniform distribution is determined from a probability density function that contains equal values along some interval between the points a and b.

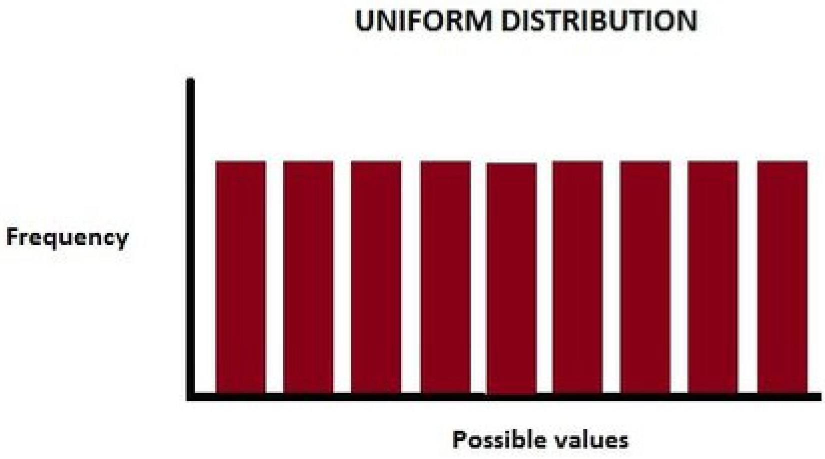
Basically, the height of the curve is the same everywhere between these two points.

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:

$$f(x) = \frac{1}{b-a}$$

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.



D] Chi-squared distribution:

There are two types of chi-square tests. Both use the chi-square statistic and distribution for different purposes:

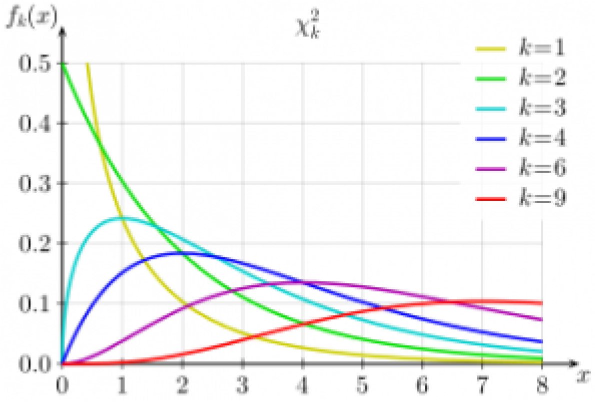
Note: We use the chi-square test in the case of testing population variance with specified value.

1] A chi-square goodness of fit test determines if sample data matches a population.
To test if the sample is coming from a population with specific distribution.

2] A chi-square test for independence compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each other.

Hypothesis formulation for chi-square goodness of fit test:

H0: The data follow a specified distribution.
Ha: The data do not follow the specified distribution.



Chi-Square goodness of fit:

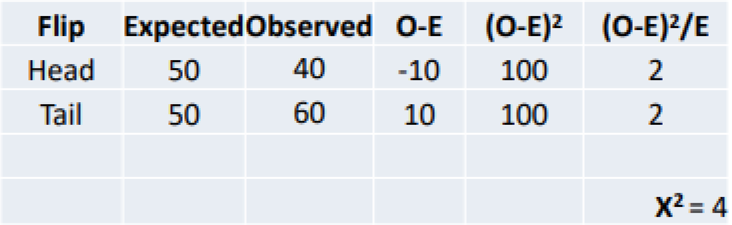
If you have a single measurement variable, you use a Chi-Square goodness of fit.
Note: In the Chi-Square test we only have a one-tailed test.

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 not biased. & Ha: Coin is 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
					X² = 4

Chi-square calculated = 4
Chi-square critical = 3.84

Conclusion: Chi-square_calc > chi-square_critical then rejects null hypothesis.

2] Chi-Square test of independence:

The Chi-Square test of independence is used to find the relationship between two discrete variables.In this case we use contingency tables.

Hypothesis formulation:

Null hypothesis : There is no relationship between the row and column variables.
Alternate hypothesis: There is a relationship. Alternate hypothesis does not tell what type of relationship exists.

The steps to calculate the chi-square value are as follows:

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. Now, write them below the observed frequencies in brackets:

	Boys	Girls	Total
Pass	17 (18.5)	20 (18.5)	37
Fail	8 (6.5)	5 (6.5)	13
Total	25	25	50

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

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

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, i.e. (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

ANOVA (Analysis of variance):

When we have to analyze means of more than two samples then we use ANOVA(Analysis of variance).It is a statistical method to compare the population means of two or more groups by **analyzing variance**.We uses F-test in Analysis of variance

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 .

Note: In this example If we use ANOVA with 95% confidence, there are 5% chances of making a mistake.

There are Two types ANOVA:

1] One-way ANOVA: In this case we have only one categorical variable(one factor)
2] Two-way ANOVA: In this case we have two variables (two factor)

One-way ANOVA is a hypothesis test in which only one categorical variable or single factor is considered. With the help of the F-test it helps us to compare the means of three or more samples.

Hypothesis Formulation in Anova:

The Null hypothesis (H0) :	All population means are same
Alternative hypothesis(Ha)	There is a difference in at least one mean.

Note: ANOVA is always one tailed test.

Two-way ANOVA is used when we need to examine two independent factors on a dependent variable.
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. Two-way ANOVA can be used to find the relationship between these dependent and independent variables.

Note: By plotting Box-plot we can see the mean of the samples & if the mean of one sample is different then reject the Null hypothesis.

ANOVA Assumptions:

0] Measurement of the dependent variable is an interval or ratio level.
1] **Random sampling:**subjects are randomly sampled for the purpose of significance testing

2] Homogeneity of variances:dependent variables should have the same variance in each category of independent variable.

F-Ratio = Mean variance between groups

—---

Mean Variance within groups

In the F-test we calculate the F ratio and compare it with the critical value and then decide on rejecting the null hypothesis or fail to reject the null hypothesis .
Variance =

$$s^2 = \frac{\Sigma(x_i - \overline{X})^2}{n-1}$$

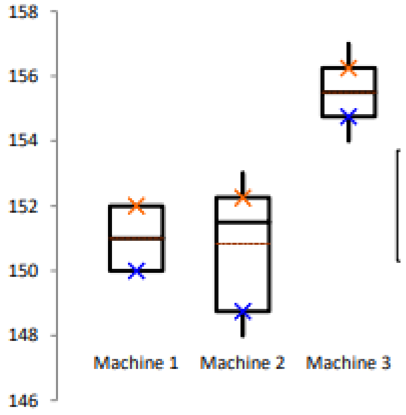
Numerator= Sum of squares
Denominator= degrees of freedom

In Anova to test whether the mean is same for all 3 machines or not we try to analyze variance within & between 3 machines.

In the below figure we clearly see mean of Machine 3 is different than other so we can directly reject null hypothesis but in complicated example we need to use Anova to make decision.

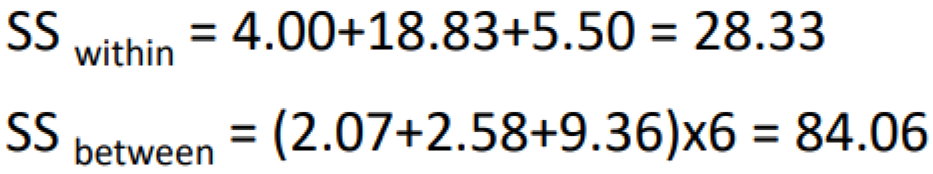
$$F = \frac{MSS_{between}}{MSS_{within}}$$

$$F = \frac{\frac{SS_{between}}{df_{between}}}{\frac{SS_{within}}{df_{within}}}$$



Example: Mean_machine1=151, Mean_machine2=150.83, Mean_machine3=155.50 check if all mean are same?(with 95% confidence)

Null Hypothesis: All three means are equal
Alternative Hypothesis: At least one mean is different



$$SS_{within} = 4.00+18.83+5.50 = 28.33$$

$$SS_{between} = (2.07+2.58+9.36)x6 = 84.06$$

Sum of square within =Variance of Machine1 + Variance of Machine2 + Variance of Machine3
Sum of square between=(mean1-Mean)^2+(mean2-Mean)^2+(mean3-Mean)^2
Mean=Mean of 3 machines(151+150+155)/3
mean1=mean of each machine.

Degrees of freedom:
df_between=3-1=2
df_within=18-1=17
df_total=df_within + df_between = 17 - 2 = 15

MSS_between(Mean sum of square) = SS_between/df_between = 84.06 / 2 = 42.03
MSS_within = SS_within / df_within = 28.33 / 15 = 1.89

F_score = MSS_between / MSS_within = = 42.03 / 1.89 = 22.24
F_critical_value = 3.68

F_score > F_critical so, At least one mean is different out of three.we can reject the null hypothesis.

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.

1. Make a Hypothesis
    1. Null hypothesis or H0:
       ” there is no difference in the conversion rate in customers receiving newsletter A and B”.
    2. Alternative Hypothesis or H0:
       Ha is- “the conversion rate of newsletter B is higher than those who receive newsletter A”

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.

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%.

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.

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