**Statistics:**

Variance of the sample is lower compared to the variance of the population in most of the cases, cause the chances of having outliers in the sample is low as the chances of having the general most occurring element is high.

Why use n-1 instead of n while calculating the sample variance,

1. **Biased sum of squared difference:**

The sum of square distances from sample mean are always lower (underestimated) compared to the sum of square distances from the population mean, cause the mean calculated in the samples is far closer to the points (major chunk of the points) than the actual mean. Meaning the sample mean is not the exact correct mean it is the biased

most probable (of similar kind) in the dataset as the most probable points will have higher probability to be in the sample in the first place, cause the extreme outlier points will have lower probability to be in the sample. So, point is sample mean is more tends towards the this highly occurring points and won’t have any outlier extreme essence in it. So, if you calculate the squared sum of the differences by this sample mean it will be smaller compared to if we calculate by population mean. Let’s take an example for calculation, suppose we have sample mean as well the population mean, now let’s pick two random points form the population and calculate the sum of squared differences from the both the means, the value of sum of squared differences from the sample mean is always lower compared, compared to the value form the population mean. So, the squared sum of the differences is biased cause of the mean in accuracies.

1. **Degree of freedom:**

So, the above reason tells you why the sample variance does is not same as the population variances and tells you it is biased and underestimated, so to reduce the bias, the denominator is reduced, DOF tells you by what amount the denominator to reduce by.

Calculating the variance with sample mean, reduces the DOF by 1 as, calculating the mean already makes the use of one value redundant as it can be calculated by mean and rest available values, reducing the independent information value to n-1.

**Understand why is it that while calculating the percentiles we do use n+1 rather than using the “n” only.**

Where to use the cumulative probability function, in cases where the probability is compared against the known threshold, like financial risk management, calculating the probability of the following below or above a certain threshold.

**Percentile:**

Percentile location/value (index form percentile number) is calculated as that proportionate position after ordering the numbers, i.e. 75th proportionate position out of all based on the order. 0.75\* N, this the index, get the number belonging to this index.

Percentile number (i.e. what percentile from the location/index or the value), number values below this value to the total number of values. The proportionate is the percentile. In case of the repetition of the value use mid rank approach i.e. add the 0.5 X (no. of times the value is repeated). Using 0.5 which means in a way that half the values are below the current value which we are counting and that is why add this second term indicating this many values below the current value while other 0.5 X n indicates the value above the current value which we are skipping.

**Covariance: C**ovariance tell you the kind of relationship between the two variables. It does not tell you the degree of relationship.

Intuition is as such.

Check for each point of variable whether it is above or below the variable average then do the same for the other variable, get the summation of the product of these two quantities. The sign indicates the kind of relationship positive or negative. Take the average like you do for the variance, with n-1 for samples and n for the population.

Each difference represents if the corresponding datapoint is above or below the average of the variable and if the two above average datapoints multiplied together they give positive value and same for the two negative values. Indicating they have correlation that is increasing one increase other, and similarly for decreasing one decreases the other. And you take the summed average to see the same thing for entire dataset.

**Covariance** does not give you the strength of the relationships because it not normalized value. The units of both variables could be changed, and the covariance could be changed. So, basically what you do is, normalized the covariance with it’s per unit variability, and that is why it is between (0-1) and (0-(-1)).

So, correlation depends on variability (SD) of the variables. If the variability is less even small positive spread across both variables (i.e. COV between variables) will tend to have high degree of overall CORR strength and vice versa).

Why use multiplication instead of summation in covariance, cause the summation won’t able to preserve the sign of each deviation (summation of differences from mean is zero for both), cause the most important thing that needs to be measured is the direction (sign) of each datapoints which describes or gives you the indication of how each datapoint moves (i.e. in positive direction or negative), as in summation the sign or direction preservation is based on the magnitude of the sum of differences of datapoint from their means respectively. E.g. suppose the first difference is 10 and second one is -3, the summation would be 7 but for the entire it would be zero and won’t make sense at all, it does not preserve the sign or direction though the relation for a single point is completely negative as one is higher than mean correspondingly the other one is decreasing with ref to its mean. But the same with multiplication can be preserved.

Correlation, which normalizes the covariance to their variability (variability of each variable) by dividing the covariance by the product of the SDs of two variables. I still do not understand the intuition of the dividing by the SDs.

Coefficient of variation: why needed and what is the intuition of it?

Variance/SD gives the idea about the dispersion in the dataset, but it is the absolute variation in the dataset. The meaning/interpretation of variation is specific or dependent on the magnitude of values. So, to have the measure of dispersion with respect to the magnitude of values or central tendency of the dataset coefficient of variation is calculated which gives the measure of the dispersion with respect to the measure of central tendency. Useful, mainly in the cases where we need to compare the variability of two datasets.

**Why the y-axis i.e. probability density in the probability density function** is not the exact probability itself, but it is the probability density or rather a probability per unit (whatever could be the unit of the x-axis) and hence the actual probability is calculated for two points only.

**Density estimation,**

**Parametric and non-parametric:**

Fundamentally the parametric means it assumes standard fixed form globally throughout with fixed parameters. (like linear regression in ml models, like normal distribution in density estimation).

Parametric model are models with parametric distribution assumption about the data, whose density function is standard formula for that function with parameters calculated from the data. For normal distribution assumption of the data parameters would be (sigma and mean).

Non-parametric estimation when the standard form is not directly assumes for the distribution of the data globally, because they cannot be assumed to be following the standard distribution, the density function for these is calculated with the help of the kernels which are standard parametric forms of the distributions only. Unlike the parametric which are defined globally for the entire dataset, in non-parametric with the help of the kernels local density is established.

The main key difference is parametric models does not grow with data, which is not the case for the non-parametric model which grows with data, like in case of the DT, KNN, SVM with RBF etc.

How to calculate the non-parametric?

Probability distributions?

* Probability density function (for continuous numerical data).
* Probability Mass Function.

Skewed and non-skewed distribution mean, mode and median distribution?

* How does it affect, very important image to remember, so firstly we know that for perfect normal distribution mean, mode and median all are overlapping. Now, supposedly we add few bigger extreme values in the distribution what happens then, distribution becomes right skewed and the mean shifts towards the right as the higher values are added, and so does the median cause now the range is increase towards higher side. And why does median is lower than mean, the answer lies in using median over mean in the presence of the outliers.
* Same thing for left skewed.
* Thing to remember, do not compare the left and right skewed straight away, compare the left skewed with the perfectly normal and same for the right skewed.

What is **kurtosis** and why it is there? first let’s understand how distributions are described.

Various shapes of the data are mapped by various distributions with its own parameters, like normal distribution is by mean and sigma, log by almost similar parameters and many more. So once mapped, with the help of the distribution’s probability density function, probability can be calculated. Whereas the kurtosis and skewness can be calculated for each distribution.

**Skewness for the perfect normal distribution is 0 and kurtosis for the same is 3.** If any of these two changes the distribution is not normal (perfect) normal distribution. To make the distribution non-normal change either of the above two. Pdf formula does not have both values included in it. Cause the condition where the pdf formula valid i.e. for perfect normal distribution these values are absolute constant. If the kurtosis is not 3 for the distribution, then the distribution becomes non-normal and the probability for the same can be estimated by any method of density estimation of non-normal data like KDE.

It’s not like each distribution/shape, or its probability density function is given by mean, sigma, skewness, kurtosis. Each distribution has its own parameters though this one can be calculated for each one.

Skewed data can be model by log normal, gamma and many other distributions,

Each distribution has different value of the kurtosis, for normal distribution this value is 3. Reason is rooted in the mathematics of it.

How to calculate the kurtosis?

* Calculate the fourth central moment just like variance (second central moment) just instead of degree 2 calculate with degree 4 or 4th power of the difference.
* Then the kurtosis would be (forth central moment) / (sigma\*\*4).
* Why central word in moment, cause the deviation is calculate from mean.
* Excess kurtosis is kurtosis - 3, but why?
* **Kurtosis for the normal distribution is 3** always, so compare relatively with normal distribution excess kurtosis is calculated which shows the relative of the kurtosis indicating whether the tail is fatter or narrower.
* Excess kurtosis greater than 0 is leptokurtic, thicker tails
* Excess kurtosis lower than 0 is platykurtic, thinner tails (plateau, i.e. platykurtic)
* Excess kurtosis equal to 0 is mesokurtic. (middle way to remember naming)

**How to know if the data distribution is normal or not?**

* QQ plot
* Statistical test like, Shapiro Wilson test, Anderson-darling test.

**Normal distribution:** e.g. height distribution, weight etc. Z represents how far the point is form the mean in terms of SD i.e. how many SDs away the point is from the mean.

**Uniform distribution:** uniform probability of all X’s e.g. drawing random number from the 1-100, drawing a card form the well shuffled pack of cards, rolling a dice and getting a number. Etc.

Used in many cases like random initialization, data augmentation, sampling, and hyperparameter tuning by defining the range of hyperparameter (or already known a range and then selecting the arbitrary random values within that range)

**Log normal distribution** will always be right skewed, the log normal distribution is exponential of the normal distribution, and the percentage increment by exp with respect to the original data point is more for higher datapoints, so the higher points would be placed far towards the right than the point compared to the lower points, which won’t shift as much towards the right.

**CLT:**

There is something called sampling distribution, where the distribution of the sample stat is plotted. Sampling distribution with mean and sum stats CLT comes into picture.

CLT is the fundamental basis because normal distribution works.

Properties of the CLT, mean of the distribution of the samples means is the equal to the mean of the population, whereas the SD of the sample means is SD (population)/sample size. (not the number of samples).

Basically, now we have two distribution one is population distribution and other one is sample means/sums distribution the mean is same for both (i.e. mean of the sample means is equal to the population mean) and SD for the sample distribution is population SD/ sample size.

This is the property use in population parameter estimation by confidence interval, statistical testing, etc.

If the sample size is not 30 then sampling distribution resembles the underlying population distribution. Meaning, the sampling distribution is normal only for normally distributed population data. There is no idea of student’s t distribution here, based of sample size and all, it only exists sample SD is used to resemble/estimate the population SD.

**Standard Error:** It is the standard deviation of the sampling distribution, i.e. SD of the sample statistic (mean) meaning how far away this sampling mean/stat from the distribution mean or population mean. It is called standard cause error because it is the SD of the sampling distribution and error because it represents the error in estimating the population parameter from the sample, i.e. error/difference of our sample mean and actual population mean which is the mean of the sampling distribution, basically standard error represents the distance of point from the mean(population). Higher the sampling size lower is the SE, cause the uncertainty associated is reduced with higher number of samples.

SE = SD/root of N. SD is the population SD when population SD is known, else it is samples SD, for multi sample test the SE is calculated accordingly with certain of aggregation.

**Confidence intervals:** Estimating the population parameter from a single sample to lie in certain range of values with certain degree of confidence.

**we cannot be confident that this very first sample’s confidence interval contains the true population parameter**-there is **no guarantee** for any individual sample’s interval.

How is CLT use in the case of the CI? You know what CLT states which tells that mean of samples mean distribution is equal to the mean of the population and SD sample mean distribution is SE which is population SD/ root of sample size.

Now is this useful in estimating CI, to make CLT we need multiple samples, and you cannot create multiple samples from data which is available then they won’t be independent samples even the very large data is available. So, **from single sample we predict the CI for population mean (parameter). How do you do so?**

* We know the sampling distribution follows the normal distribution, and we have single sample mean i.e. a single datapoint form this normal distribution of the sample means if there were multiple samples drawn obviously.
* Now from this single datapoint of the normal distribution whose mean is equal to population mean and SD is SD of the population / root of sample is size, we need to estimate the mean given the sample size and SD of the population.
* We know the sampling distribution follows the normal distribution, and we have single sample mean i.e. a single datapoint form this normal distribution of the sample means if there were multiple samples drawn obviously.
* Now from this single datapoint of the normal distribution whose mean is equal to population mean and SD is SD of the population / root of sample is size, we need to estimate the mean given the sample size and SD of the population.
* point is if the sample size is more than 30 or large in general then the underlying distribution of the population does not matter, but if the sample size is small then in that case it handy to have underlying distribution as normal.
* A 95% confidence interval is centred at the sample mean, not the population mean, so it does not always cover the central 95% of the sampling distribution. Instead, 95% refers to the long-run proportion of such intervals that will contain the true population mean
* A confidence interval (CI) estimates a range around the sample mean that, based on the chosen confidence level (like 95%), is likely to contain the true population mean-not a certain proportion of the sample data, but a certain proportion of possible sample means if the sampling were repeated many times.

**Where and when to use Z-Procedure?**

I say you want to estimate the mean of the national level test score for this year based of sample score of 100 students where 100000 students gave the exam, though the SD of the exam for last 50 year is available, can I say this is the example where Z procedure is used

* If you have a **known population standard deviation ($\sigma$)**-either from historical data or from a very large, reliable sample-you can use the **Z-procedure** for confidence interval estimation, provided your sample size is reasonably large (usually n≥30*n*≥30).

**Where and when t-procedure is use:**

* When the SD of the population is not known, the SD of the samples is used, which increases the uncertainty which ultimately changes the distribution making the distribution as students t-distribution with fatter tails to accommodate the uncertainty involved due to use of sample SD instead of population SD. When the sample size
* the point is if the SD of the population is not known then you always use t-procedure just the fact that if the sample size is high, this means the t test is more reliable cause t distribution for large sample gets closer to normal distribution, is it the case
* **Student’s t distribution**, symmetrical bell-shaped curve with thicker tells which accounts for more extreme scenarios, used in case when the population SD is unknown, and sample size is small. Students-t distribution is constructed as distribution of the sample means of the samples when the SD of the population is not known. It is the pdf of the t-statistics for DOF.
* Cause the SE of the sample distribution is calculated from the sample SD instead of population SD that is why the distribution does not remain normal and becomes the students T distribution, which accounts for the more uncertainty which is involved due to use of SD of the sample.
* For higher sample sizes, the t-distribution becomes increasingly like the normal distribution.
* The t-distribution is a normal-like distribution with higher kurtosis (heavier tails), especially noticeable with small sample sizes.
* So, theoretically where we use t distribution, i.e. we assume the sample statistic follows student’s t distribution instead ideally it should follow perfect normal in if the sample size is high. So, the reason for using the t distribution is only to compensate the biased or randomness involved because we are not able to correctly get the SD of the sampling distribution?

Things to do in details after,

* PNC
* CONVOLUTION
* CLT and NORMAL INTUITION.

**Hypothesis Testing:**

So, basically, all we do is, we consider the null hypothesis, i.e. state of no change as true and then try to find the evidence whether it is false or not, if we are not able to find evidence against it, then we stick with our presume opinion of null hypothesis being True and reject the proposed change i.e. alternate hypothesis, and vice versa. There is nothing as proving something It’s all about accepting null or not based off the evidence from the data.

You use the location of your sample statistic in the sampling distribution to determine how likely it is to occur by chance. If it’s very unlikely, you reject the null hypothesis. This is the essence of statistical hypothesis testing.

The core intuition of the hypothesis testing is, you try not accepting the alternate hypothesis without evidence, so try to kind of reject it with the data from the sample while finding evidence against null hypothesis, only if we you find evidence against the null hypothesis you accept the alternate hypothesis.

Why there are two hypotheses? usually the alternate hypothesis and null hypothesis are kind of complementary to each other, they are fully complementary sometimes, but not all the time. Wherever they are totally complementary then in that case only one is required.

**so, whichever we accept or reject we can't surely say the one is guaranteed correct or not?**

Certainly! Here’s your paragraph, revised for clarity and accuracy:

In hypothesis testing, we start with a null hypothesis, which we want to test for evidence against. We do this by checking where the sample statistic (such as the sample mean) falls within the sampling distribution under the assumption that the null hypothesis is true. If the sample statistic falls beyond a certain critical region (determined by our chosen significance level, such as 95%), this is considered evidence against the null hypothesis. For example, in a two-sided z-test where the null hypothesis is that the population mean equals 50, and the alternative hypothesis is that the population mean is not equal to 50, we calculate the z-score of the sample mean. We then determine the boundaries (critical values) that contain 95% of the sampling distribution around the mean (typically z = ±1.96 for a 95% confidence level). If the sample mean falls within this central region, we do not reject the null hypothesis. If it falls outside this region (in the outer 5%), we reject the null hypothesis, concluding that there is significant evidence that the population mean is not equal to 50.

e.g. in insurance

**Testing Average Claim Amounts**

* **Scenario:** An insurance company originally set policy rates assuming the average claim amount is $1,800. They are concerned the true mean may now be higher, which could impact profitability.
* **Hypotheses:**
  + Null hypothesis (H₀): Mean claim amount ≤ $1,800
  + Alternative hypothesis (H₁): Mean claim amount > $1,800
* **Process:** The company samples 40 claims and finds a sample mean of $1,950. Using a t-test, they check if this increase is statistically significant. In the example, the test statistic did not exceed the critical value, so they failed to reject H₀, meaning there was not enough evidence to conclude the mean claim amount had increased. In this case the SD of the sample is used.

**Comparing Claim Amounts Between Groups**

**Example:**  
An insurer wants to know if smokers file higher medical claims than non-smokers.

* **Null hypothesis (H₀):** Mean claim amount for smokers = mean claim amount for non-smokers
* **Alternative hypothesis (H₁):** Mean claim amount for smokers > mean claim amount for non-smokers

A t-test is used to compare the average claims of the two groups. If the p-value is below the significance threshold (e.g., 0.05), the insurer concludes that smokers have higher claims

In a **Type I error,** the null hypothesis is rejected more often than it should be (so the alternative hypothesis is incorrectly selected more frequently), whereas in a **Type II error**, the alternative hypothesis is not selected as often as it should be (because the null hypothesis is incorrectly retained.

**Generally, Type 1 errors are considered more serious in many fields.**

If a new cancer medicine is falsely found effective due to a Type 1 error, it gets approved and prescribed despite not actually working. Patients may rely on this ineffective drug, risking their health and missing out on real treatments. This makes a high Type 1 error rate extremely dangerous in medical research. and in case of type 2 error, it mainly represent the missed opportunity, but this is very specific the problem how test is designed.

case where accepting the wrong alternate is highly costly then type 1 is more detrimental and where accepting the wrong is null is more costly then type 2 is more severe and detrimental.

One side and two side tests.

The **p-value** is the probability of obtaining a result as extreme as, or more extreme than, what you observed in your data, **assuming the null hypothesis is true.**

Not the probability of rejecting the null hypothesis. The p-value is the probability that measures the strength of the evidence against the null hypothesis

The **significance level**, often denoted by the Greek letter alpha (α), is the threshold set by the researcher before conducting a statistical test to determine how strong the evidence must be to reject the null hypothesis. It represents the **maximum probability of making a Type I error**-that is, the risk of incorrectly rejecting the null hypothesis when it is true.2

For example, a significance level of 0.05 means you are willing to accept a 5% chance of concluding that an effect exists when it does not. If the p-value from your statistical test is less than or equal to the significance level, you reject the null hypothesis and consider the result statistically significant.

**The test stat follows the null hypothesis distribution, and if it lies far away i.e. with lower probabilities meaning chances of it lying in the null distribution is less.**

Different types of the hypothesis test?

1. **What is a Z-Test?**

A **z-test** is a statistical hypothesis test used to determine whether there is a significant difference between sample and population means, or between the means of two independent samples, **when the population variance (or standard deviation) is known,** and the sample size is large (typically n ≥ 30).

**Z-tests are appropriate when:**

* The data are approximately normally distributed (or the sample size is large enough for the Central Limit Theorem to apply).
* The population standard deviation is known.
* The data points are independent and randomly selected.

**How Does a Z-Test Work?**

State the Hypotheses:

**Null hypothesis (H₀):** No difference or effect (e.g., the sample mean equals the population mean).

**Alternative hypothesis (H₁):** There is a difference (e.g., the sample mean does not equal the population mean).

**Set the Significance Level (α):**

Commonly 0.05 (5%).

**Calculate the Z-Statistic (Z-Score):**  
The z-score measures how many standard deviations the sample mean is from the population mean:

* Z=Xˉ−μ0σ/n*Z*=*σ*/*nX*ˉ−*μ*0
* Where:
* Xˉ*X*ˉ = sample mean
* μ0*μ*0 = population mean under the null hypothesis
* σ*σ* = population standard deviation
* n= sample size

**Compare to the Critical Value or Find the p-value:**

* For a two-tailed test at α = 0.05, the critical z-values are ±1.9
* If the calculated z-score is more extreme than the critical value, **reject the null hypothesis**.
* Alternatively, calculate the p-value and compare it to α.

**Draw a Conclusion:**

* If |z| > critical value or p-value < α, reject H₀.
* Otherwise, do not reject H₀
* **Types of Z-Tests**
* **One-sample z-test:** Compares a sample mean to a known population mean.
* **Two-sample z-test:** Compares means from two independent samples, same as the two sample test, just the difference is it uses population SD instead of sample SD. E.g. comparing the mean IQs of the two cites when the SD of the two cities population is known.

**Example**

* Suppose a company claims the average salary of its employees is $50,000 (σ = $5,000). You take a random sample of 40 employees and find an average salary of $52,000. You can use a z-test to determine if this difference is statistically significant
* **In summary:**  
  A z-test is a hypothesis test for comparing means (or proportions) when the population standard deviation is known, and the sample size is large. It uses the z-score to measure how far your sample statistic is from the hypothesized population value, helping you decide whether to reject the null hypothesis

1. **T -tests:**

**One-sample t-test:** Compare the mean of a single group to a known value or population mean

**Independent two-sample t-test:** Compare the means of two independent groups

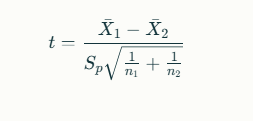
**Paired (within-subjects) t-test:** Compare means from the same group at two different times or under two conditions (e.g., before and after an intervention).

One sample t-test: The intuition of the single sample test is same as the z-test,

e.g. A nutritionist claims that the average daily protein intake for adults in a certain city is 50 grams. To test this claim, a researcher randomly selects 15 adults and finds their average protein intake is 54 grams with a standard deviation of 6 grams. T statistic (T score) is calculated by the given data as (54-50)/SE, where SE is SD of sample upon root of sample size.

e.g. 2. An insurance company historically processes claims at an average cost of $60, but with rising expenses, they want to test if this has changed. They sample 20 recent claims, finding a mean cost of $65 and a standard deviation of $8. Using a one-sample t-test, they calculate a t-value of approximately 2.80 (t=65−608/20*t*=8/2065−60), which exceeds the critical value of 2.093 for 19 degrees of freedom at a 0.05 significance level. This result leads them to reject the null hypothesis and conclude that the average claim processing cost has significantly increased from $60.

**Two sample tests:**



*df* =*n*1+*n*2−2

**The purpose of the independent two-sample t-test is to check whether two samples, drawn from two independent populations that are assumed to have equal standard deviations (SD), have the same mean or not.**

Intuition:

*The independent two-sample t-test helps you answer a simple but important question:****Are the average values (means) of two separate groups different, or could the observed difference be due to random chance?***

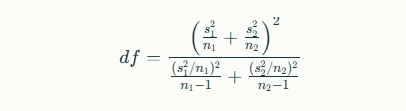
*The t-test measures how many “standard errors” apart the two means are. The bigger this number, the more confident you can be that the difference is not just due to random variation.*

**You’re checking whether the difference between the two-group means is large relative to the variability (standard deviation) within the groups.**

**e.g.** The company calculates the average claim amount for each group and applies a two-sample t-test to compare the means. If the t-test shows that new drivers have a statistically significantly higher average claim amount (with a low p-value), the insurer can confidently conclude that new drivers represent a higher risk. This evidence justifies charging higher premiums to new drivers compared to experienced ones, ensuring that the pricing structure accurately reflects the underlying risk

**Welch’s t-test**

t-stat is calculated in similar manner just the difference is how is df calculated and it is mentioned below.



Intuition of CLT in Two-Sample Tests

* **As sample sizes grow, the means of each sample (from their respective populations) become approximately normally distributed, regardless of the original population distributions.**
* **Because both sample means are (approximately) normal, the difference between the two sample means also follows a normal distribution when the samples are large.**
* This is true even if the underlying populations are not normal, if the samples are sufficiently large and independent.

Why This Matters

* The normality of the difference in sample means (thanks to the CLT) allows us to use t-tests and calculate p-values, making statistical inference possible and reliable in real-world scenarios.

What You’re Doing in a Two-Sample Test

* You **calculate the difference between your two-sample means**.
* The **Central Limit Theorem** tells you that, for large enough samples, the **sampling distribution of the difference between means** will be (approximately) normal.
* You then **ask: “Where does my observed difference fall within this theoretical distribution?”**

Interpreting the Result

* If your observed difference is **close to the centre** of the sampling distribution (near zero, assuming the null hypothesis that the means are equal), it’s likely just due to random chance.
* If your observed difference is **far in the tails** (very unlikely under the null hypothesis), you have evidence that the population means are different.

e.g. You measure the grams of protein in two different brands of energy bars. Each brand represents an independent group. A two-sample t-test helps you assess whether the average protein content differs between the two brands

**paired sample t – test:**

Students take a practice test and then, after a study session, take a final test. You use a paired t-test to see if the study session significantly improved their scores, since each student’s scores are paired (pre- and post-study)

Students take a practice test and then, after a study session, take a final test. You use a paired t-test to see if the study session significantly improved their scores, since each student’s scores are paired (pre- and post-study)

The key **difference between a two-sample t-test and a paired sample t-test** is that the two-sample t-test compares the means of two independent, unrelated groups (such as different people in each group), while the paired sample t-test compares the means of two related groups where each data point in one group is matched to a specific data point in the other group (such as measurements from the same person before and after a treatment); essentially, the two-sample t-test assesses differences between groups, whereas the paired t-test evaluates changes within matched pairs.

**Chi square test:**

**Chi square distribution:**

So, chi square is just the naturally occurring phenomenon or distribution occurs when the random variable is taken from the standard normal distribution and square them then sum them and you do it for n time, then you get the chi square distribution, as DOF increases this distribution becomes normal.

In chi square test location of sample chi square stat is found out and checked where does it lie in sampling chi square distribution?

Why chi square is non-parametric?

*O follows a binomial distribution, and for large n, the binomial is well-approximated by a normal distribution.  
This is a key reason why the chi-square test works.*

*So, the O is the normally distributed for large sample size.*

same fundamental is use while changing the normal distribution to standard normal as well, where the mean is E and variance is also E cause for high n in binomial with low probability the variance becomes E, and mean is Expected or average of outcome which is also E.

and this is how we get the normal sample stat, which is follows chi square distribution, cause this sample stat is calculated by squaring the independent random variables (categories value in actual test) which are from the standard normal distribution and how this one value which is squared is standard normal distribution is explained above.

and this is how we get the normal sample stat, which is follows chi square distribution, cause this sample stat is calculated by squaring the independent random variables (categories value in actual test) which are from the standard normal distribution and how this one value which is squared is standard normal distribution is explained above.

**Why do we need CLT?**

For large sample sizes, observed counts (O*O*) from a binomial distribution can be approximated by a normal distribution with mean and variance E=np, thanks to the Central Limit Theorem (CLT). This allows the standardized difference to be treated as approximately standard normal. When these standardized differences are squared and summed across categories, the result follows a chi-square distribution, which forms the basis of the chi-square test. The CLT is thus essential for justifying the use of normal and chi-square approximations in categorical data analysis.

In one accept of CLT is, means of samples follow normal distribution, which is handy in hypothesis testing and CI estimation, similarly other aspect is, if the squares of the **standard normal distributed** random variables are taken, they follow chi quare which becomes normal for large n or DOF.

* **CLT:** Means (or sums) become normal as sample size increases.
* **Chi-square:** Sums of squared standard normal become chi-square distributed, and for large k*k*, this distribution itself becomes approximately normal.

expected distribution meaning, what are the expected counts of each category

**Goodness of fit:** The goodness of fit test is a statistical hypothesis test used to determine how well your observed data match the expected distribution specified by a theoretical model. Most commonly, the chi-square goodness of fit test compares the observed frequencies in each category of a categorical variable to the frequencies you would expect if the data followed a particular distribution (such as equal proportions or a known probability distribution). If the observed and expected values are close, the model is considered a good fit; if they differ significantly, the model may not fit the data well.

Check notes for more revision.

**Example**

**Expected distribution:** Each number (1–6) should appear 10 times (if the die is fair).

Observed data: You count how many times each number appears.

The test checks if the observed counts are close enough to 10 for each number to conclude the die is fair

**Claim severity categories:**

An insurer expects that, based on historical data, 20% of claims will be low severity, 30% moderate, and 50% high severity. After a year, the actual observed proportions are different. The chi-square goodness of fit test can be used to check if these observed frequencies significantly diverge from expectations. If so, it may indicate a change in risk profile or require adjustment of reserves and pricing strategies

**Claim frequency distribution:**

Suppose an insurance company models the number of auto accident per policyholder per year using a Poisson distribution. After collecting data from 700 drivers, the company compares the observed number of claims per driver to what the Poisson model predicts. The chi-square goodness of fit test helps determine if the observed distribution of claim frequencies fits the theoretical Poisson distribution, validating or questioning the suitability of the model.

**Fraud detection:**

Insurers may use the chi-square test to detect fraud by comparing the distribution of certain claim attributes (such as claim amounts, types, or timing) to what is expected under normal, non-fraudulent circumstances. Significant deviations flagged by the test can prompt more detailed investigations into potentially fraudulent claims.

**Test for independence:** The chi-square test for independence is a statistical test used to determine whether there is a significant association (relationship) between two categorical variables. It assesses whether the distribution of one variable is independent of the other by comparing the observed frequencies in a contingency table to the frequencies expected if the variables were independent.

The statistical intuition is same as above for the tests, whereas the actual working with the formulas is in the notes.

e.g. association between the education and exercise.

Suppose an insurance company wants to know if the **type of claim filed** (e.g., accident, theft, natural disaster) is **independent of the policyholder’s age group** (e.g., under 30, 30–60, over 60).

* **Data:** The company collects a contingency table showing the number of each claim type filed by policyholders in each age group.
* **Null Hypothesis (H0):** Claim type and age group are independent (no association).

in above case if the test stat falls out of null region, then the age group and claim filled are dependent

**ANNOVA:**

**F distribution:**

* Ratio of two independent chi-square distributed variables, each divided by its respective degrees of freedom, represents the ratio of two variance estimates, and it follows f distribution.
* What do mean by variance estimate in this case, variance estimates meaning the variance of the sample from to which is this chi square random variable corresponds to, meaning it is the ratio of the variance of the two samples whose sum of squared of all samples represents the RV in chi square distribution.
* The term "chi-square independent variable per degree of freedom" refers to the average value of the sum of squared independent standard normal variables per degree of freedom, just like the mean of the sum of RV’s.
* For the F-distribution and F-test, samples must come from two independent, normally distributed populations. Sample variances from these samples, when appropriately scaled, each follow a chi-square distribution. The ratio of these two independent chi-square variables (divided by their degrees of freedom) forms the F-statistic. The F-statistic is used to compare the variances of the two populations-if the populations are normal, this ratio follows the F-distribution.
* Parameters are dof1, dof2, a curve is defined for combinations of dof1, dof2. Then based on f stat and this DOFs combination in cumulative f distribution finds out the probability.
* F distribution is used in:
* Comparing means of three or more variables (Analysis of variance)
* Comparing variance(F-test):
* Regression analysis to test the overall significance.

**One-way ANOVA:** One-way ANOVA (one-way analysis of variance) is a statistical method used to test whether there are significant differences between the means of three or more independent groups.

e.g. Suppose a researcher wants to know if three different teaching methods (Lecture, Workshop, Online) result in different average exam scores for students

Null hypothesis (*H*0): The mean exam scores are equal across all teaching methods.

Alternative hypothesis (*Ha*): At least one teaching method has a different mean exam score.

E.g. An insurance company wants to determine whether the average claim amounts differ significantly across four geographic regions (North, South, East, West).

How is this f distribution used in the Anova?

* F\_stat in Anova is (SSB/ ) /DF (SSW / DF),
* Where the numerator SSB does not follow chi square, but the SSB/ SD (population) follows the chi square distribution, likewise for the SSW and population SD (population is cancelled out in here.
* So, the aim is to find out the ratio of variance ratio of SSB (how much group mean differs from individual mean) and SSW (How much data points differs within each group with respect to group mean) i.e. ratio of variance between groups and variance within groups.
* The intuition of the SSB is how far away the individual mean lies compared to the population mean, if these individual means are close to population meaning these itself would be away from each other.  larger SSB suggests that the group means are more spread out from the grand mean, indicating potential significant differences among groups. Whereas SSW indicates the variability within each group, indicating the spread of each groups data. if the variance within a group is low, the mean of that group is generally a more reliable and representative estimate of the group's true value.
* So, in Anova what happens is, you try to check the variability of each group mean with population mean with respect to individual average variability of each mean.

**F- test for variance check:**

The **F-test** is a statistical test used to compare variances. It is most applied to determine whether the variances of two samples (or populations) are significantly different by calculating a statistic (the F value) as the ratio of the two sample variances. This F value is then compared to a critical value from the F-distribution, considering the degrees of freedom and significance level, to decide whether to reject the null hypothesis that the variances are equal.

The F-test assumes that samples are independent, drawn from normally distributed populations, and (for some tests) have equal variances.

The further above 1 your F-ratio is (for given degrees of freedom), the more likely it is to be considered “extreme,” and the stronger the evidence that the two population variances are different.

Usually while calculating the F stat, the lower variance is kept at the bottom to get the f stat greater than 1 and how big is this ratio than 1 for given DOFs determine the probability which indicates probability for existence of this difference in variance when sample are randomly chosen from respective population.

**E.g.**

- Manufacturing company compares the samples for variance, if the two batches have same variance or not? though the mean could be same, but the variance is very important to understand the consistency of the sample quality.

- Two investment portfolios have the same average return, but one is much more volatile than the other.

- An insurance company is evaluating two types of auto insurance policies. Both policy types have the same average (mean) claim amount per year-say, $2,000. However, for Policy A, the claim amounts are tightly clustered around the mean (low variance), while for Policy B, claim amounts are highly variable-some years have very high claims, others have very low (high variance).

**How to check the normality?**

* Shapiro wilk test.
* Kolmogorov - Smirnov Test.
* Anderson-Darling Test.
* D’Agostino-Pearson Omnibus Test.

Visualization methods.

* Hist plot
* QQ plot
* Probability-Probability plot.

**Shapiro wilk test:** So, Shapiro wilk test tells you if the underlying distribution from which the sample is drawn is it the normal distribution or not?

* The test stat here does not follow any distribution (sampling distribution) instead a test stat is calculated by analytical formula and the corresponding probability is found by monte Carlo simulation.
* Once the Shapiro-Wilk test statistic is calculated analytically, the corresponding probability (p-value) is determined using results from Monte Carlo simulations, since W does not follow a standard named distribution.

**Monte Carlo simulation** uses random sampling and repeated calculations to estimate the range and probability of possible outcomes in uncertain or complex systems.