Statistics : Key concepts

Personal Notes – Bhadri Vaidhyanathan

Contents

Key Terms:

Outcome

One of the results in an experiment. Eg: Heads in coin toss is an outcome

Expected Value:

The **expected value (EV)** comes from Probability Subject (Mean is Stats) of a random variable is the long-term average of its outcomes when an experiment is repeated many times. The term **expected value** reflects that it is the "average" outcome you would expect *in the long run* if the process were repeated infinitely many times.

Eg: The probability of heads in coin toss is 0.5 but if you toss coin 10 times it might not be 0.5 but if you do it 1000 or more times the value will tend towards 0.5 hence 0.5 is called the expected value.

Entropy

Entropy is the measure of surprise which are about all other events except the one we are calculating probability

Covariance and Correlation

Covariance: Measures how two variables move together. A **positive** value means they increase together, and a **negative** value means one increases while the other decreases.

Correlation: A standardized version of covariance that ranges from **-1 to 1**, making it easier to interpret.

Example: Employee Training Hours vs. Performance Score

Let's say we have data for five employees:

Employee	Training Hours (X)	Performance Score (Y)
A	5	60
B	10	75
C	15	85
D	20	95
E	25	100

Step 1: Compute Covariance

Using the formula:

$Cov(X,Y)= \sum(Xi-\bar{X})(Yi-\bar{Y}) / n-1$

If we calculate, we get **Cov(X,Y) = 87.5**, indicating that training hours and performance score increase together.

Step 2: Compute Correlation

$r=Cov(X,Y) / \sigma X*\sigma Y$

After standardizing, we get **r = 0.98**, meaning a **very strong positive relationship** between training hours and performance.

So, while **covariance gives a raw measure**, **correlation (0.98) tells us the strength of the relationship on a normalized scale**.

Confounding Variables

A **confounding variable** is an **external factor** that affects both the independent and dependent variables, making it difficult to determine the true cause-and-effect relationship.

Eg: Hypothesis: Employees who work remotely have **higher productivity** than those in-office.

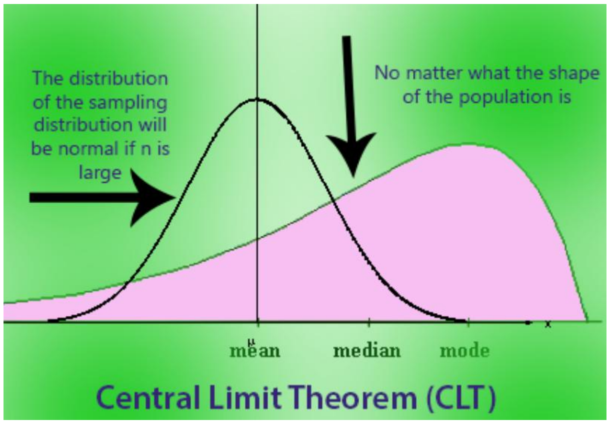
- **Confounding Variable: Job Role**
- If remote employees are primarily **software engineers** (who require deep focus), while in-office employees are **customer support reps** (who deal with unpredictable calls), then **job role**, not remote work, may be driving the difference in productivity.

Important Laws

Law of Large numbers:

The **Law of Large Numbers (LLN)** states that as the size of a sample increases, the sample mean will get closer to the true population mean. As the number of trials increase, the actual value of a probability will converge to the expected means like say head-toss is a 50:50 probability when in reality it might not be exactly 50:50 and in only few trials it could be worse like 75:25

Central Limit Theorem:


The distribution of the sampling distribution will be normal if n is large
No matter what the shape of the population is
μ
mean
median
mode
x
Central Limit Theorem (CLT)

- If you take many random samples from any population (regardless of its shape or distribution like Poisson or Binomial or Exponential) **<u>WITH REPLACEMENT,</u>** and calculate the mean of each sample…
- The distribution of those sample means will look like a **normal distribution** (bell-shaped curve)…
- **As long as the sample size is large enough** (usually n≥30).
- What if there is no replacement: In the **CLT**, the assumption is that the sampling is done **with replacement**, which ensures independence between samples. If sampling is done **without replacement**, the independence assumption is violated because the population changes as values are removed.

(Basic idea behind bootstrapping with replacement)

No matter how weird or skewed the original population is, when you take large enough samples and look at their means, those means will follow a normal distribution. The CLT is why we can use normal distribution-based methods so often in statistics!

Why Is the CLT Important?

1. **Simplifies analysis:** Even if the population is not normal, we can still make inferences using normal distribution tools if we use sample means.
2. **Foundation for hypothesis testing:** Many statistical tests assume normality of the sample mean, which the CLT ensures.

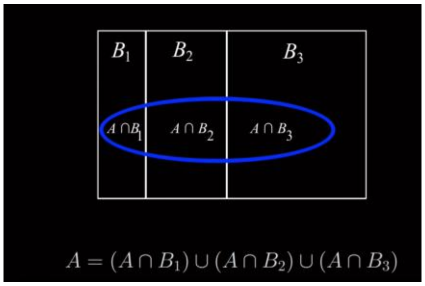
Law of Total Probability

The **Law of Total Probability** is a fundamental rule in probability that helps us calculate the probability of an event by breaking it into smaller, simpler parts based on a partition of the sample space.

B1, B2, B3 are mutually exclusive events or domains eg: 3 different factories. (say we are going to look at the defect rate of each factory. Here B1,2,3 are mutually exclusive and an event either occurred in B1, 2 or 3 and never both). A is a common event (Lets say defect rate) that happens separately in B1, B2 and B3 sample space: A is a subset of all 3. then if P(A) in B1 space, P(A) in B2 space and P(A) in B3 space is known separately, to get overall P(A) you can just add them together after multiplying with the appropriate $P(B_x)$ (conditional probability formula).

$$P(A) = P(A)_{B1} + P(A)_{B2} + P(A)_{B3}$$

$P(A)_{B1}$, $P(A)_{B2}$ and $P(A)_{B3}$ are actually arrived at using conditional probability formula like $P(A)_{B1} = P(B_1) * P(A|B_1)$.



https://www.youtube.com/watch?v=7t9jyikrG7w

Simple Example: Scenario

Suppose there are three factories B1,B2,B3 that produce the same type of lightbulbs. The probabilities that a randomly chosen bulb comes from each factory are:

- P(B1)=0.4 = 0.4 (40% from Factory 1),
- P(B2)=0.35 = 0.35 (35% from Factory 2),
- P(B3)=0.25) = 0.25 (25% from Factory 3).

The probability that a bulb is defective (A) depends on the factory:

- P(A|B1)=0.02 (2% defective from Factory 1),
- P(A|B2)=0.03 (3% defective from Factory 2),
- P(A|B3)=0.05 (5% defective from Factory 3).

Using the Law of Total Probability:

P(A) = aka P(A∩B1) + P(A∩B1) + P(A∩B1)

= P(A|B1)*P(B1)   +   P(A|B2)*P(B2)   +   P(A|B3)*P(B3)

[Remember: P(X∩Y) = P(X|Y) * P(Y) = Conditional Prob formula. This is applied above]

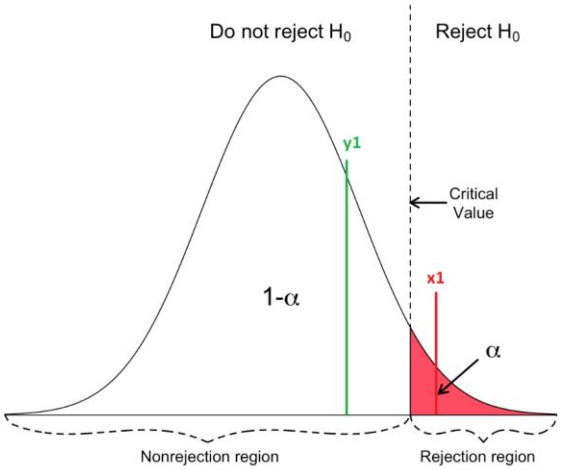
Substitute the values:

P(A) = (0.02)(0.4)+(0.03)(0.35)+(0.05)(0.25)
P(A) = 0.008 + 0.0105 + 0.0125 = 0.031

So, the total probability that a bulb is defective is **3.1%**.

Hypothesis Testing

In layman terms:
Hypothesis testing is a framework to check if a value x is close enough to population mean or too faraway. This is done by assuming the population is in a certain distribution (like normal distribution) with some mean μ and sd σ. Knowing these values, x, μ and σ and other facts, one can find the p value/probability (pink area below) of the value x. If that p value is less than 5% then it implies it is far (many σs away from mean) away from population mean.
Note: pvalue looked up in charts gives the area under curve and not a value for one point.



Based on p value, you can tell if value its falls in the critical region(too faraway) or acceptance region(close to pop. mean). The threshold is the alpha α value usually 0.05. Hence p value is less then 0.05 then it means accept alternate Hyp and reject null. If p value is above 0.05 then null hyp. is not rejected.

Take sample, using the numbers in sample make a prediction on population.

- When sample is taken and hypothesis test is done, one of the purpose is to find the probability that the difference found (alternate hyp) is large (95%) enough thus allowing us to confidently accept the alternate hyp.
- Statistically, we find the p-value for the sample in question. The resulting p value simply tells if the sample is part of null hypotheses (large p value) distribution or not part of null hypothesis distribution (small p value).

Hypothesis Testing within Linear Regression

Regression analysis is one of the most commonly used tools in data science for predicting outcomes and uncovering relationships between variables. Many aspects of regression involve **statistical hypothesis testing**:

  - Testing if the slope of a predictor variable is significantly different from zero.
  - Comparing models or determining if the inclusion of certain variables improves the model.
- **Key Tests within Regression**:
  - **t-Test on Coefficients**:
    Tests whether a specific coefficient (β) significantly contributes to the model. Null hypothesis: β=0(no effect).
  - **F-Test for Model Fit**:
    Tests whether the overall model explains significantly more variance than a baseline (e.g., mean-only model).
- **Example**:In linear regression, you might predict sales from advertising spend. You test if the slope for "advertising spend" is significantly different from 0 (does advertising matter?).
- **Relation to Hypothesis Testing**:Regression incorporates hypothesis testing to validate its coefficients, evaluate goodness-of-fit, and make inferences about relationships.

Binomial Test – Hypothesis testing in A/B testing

The Binomial Test evaluates whether the observed proportion in a binary outcome (e.g., success/failure, yes/no) is significantly different from a specified value. It often appears in A/B testing or experiments involving categorical outcomes.

- **Example**:
  - Testing if the success rate of a new website design is higher than 50%.
  - The **null hypothesis** might state: "The success rate is 50%."
  - You calculate the probability of observing your result (or more extreme) under this null.
- **Relation to Hypothesis Testing**:

- Like other tests, it evaluates a null hypothesis with a p-value.
- It's conceptually similar to other proportion-based tests like the **z-test for proportions** but tailored for small samples or exact probabilities.

Disadvantages of Hypothesis testing approach (against Machine Learning):

Eg: We want to open a mikshake shop and want to know if chocolate or vanilla shake is more popular. We did a survey with 1 million people and hypothesis testing approach showed that chocolate was more popular so we went with it.

What was apparent is that chocolate was more popular but 51% more popular and vanilla was 49% popular which practically means that the shop should sell both flavors since they difference is small. This is not apparent in Hypothesis testing approach hence Machine learning approach (like linear and logistic regression) is better.

Hypothesis testing is binary as in yes or no and doesn't provide more information or data while ML algos provide more information about the relation between the variables in question.

Mitigating errors in Hypothesis testing:

If 0.05 p value or 5% p value is used, then 1 in 20 is wrong. Our test could have been that 1 in 20 and thus giving us wrong conclusions. How do we mitigate this? One approach is called **Bonferroni** correction.

When performing more than one hypothesis test, your type I error compounds. In order to correct for this, a common technique is called the **Bonferroni** correction. This correction is **very conservative**, but says that your new type I error rate should be the error rate you actually want divided by the number of tests you are performing. Therefore, if you would like to hold a type I error rate of 1% for each of 20 hypothesis tests, the **Bonferroni** corrected rate would be 0.01/20 = 0.0005. This would be the new rate you should use as your comparison to the p-value for each of the 20 tests to make your decision.

p value

https://towardsdatascience.com/a-simple-interpretation-of-p-values-34db3777d907

- **p-value = "How surprising is my data under H0?"**
- **Small p-value (<α<\alpha<α): Surprising → Reject H0.**

- **Large p-value (>α>\alpha>α): Not surprising → Fail to reject H0.**
- **It's a measure of consistency with the null hyp., not direct evidence for the alt. hyp.**

- **P-value theory seems roundabout because it's built on the logic of indirect inference:**
- **Assume H0 is true → Check how well observed data aligns with H0.**
- **Rarity is established / assessed relative to a reference distribution (from H0) we ourselves select, not directly measured nor implied to be accurate.**

Alpha Threshold (α)

- The threshold α (commonly 0.05) determines whether to reject H0.
  - p < α (.05) : Evidence against H0 reject the null.
  - p ≥ α (.05) : Not enough evidence, fail to reject H0.
- α=0.05 corresponds to ~2 standard deviations in a normal distribution (95% confidence).

p value is the area under the appropriate probability distribution curve of the statistic (which is assumed to be representative of the statistic's OG population by an expert) being observed. The area includes other events which are equally rare and more rarer. If pvalue is high then null hypothesis does not get rejected a while alt.hypothesis is rejected and that the event being observed is part of the general population and not rare and different from the other events by referring to how close it is to the mean of the population.

- **p-value (p)**: Probability of obtaining a result that are equally extreme to or more extreme than was observed in the data.

p value = p(rare event) + p(equally rare event) + p(of all more rare events)

"Low pvalue means "rare" and so accepts alt hypothesis (new treatment) rejects null hypothesis (status quo)"

p value is the probability that random chance generated the data(1), or **something else equal(2)** or **rarer(3)**. pvalue is the probability% of an item and anything worser under the normal curve (or other distribution as used).

- defn: **p value is the probability that random chance generated the data** , or something else equal or rarer.
  - Lets split that defn into 3 and with example of getting two heads in a row.

p value is

- **Part1) the probability that random chance generated the data:**

E = TT, TH, HT,HH

P(HH) = ¼ = 0.25 is

  - Part2) **something else equal**

Other possible outcomes with same prob. is P(TT) which is 0.25

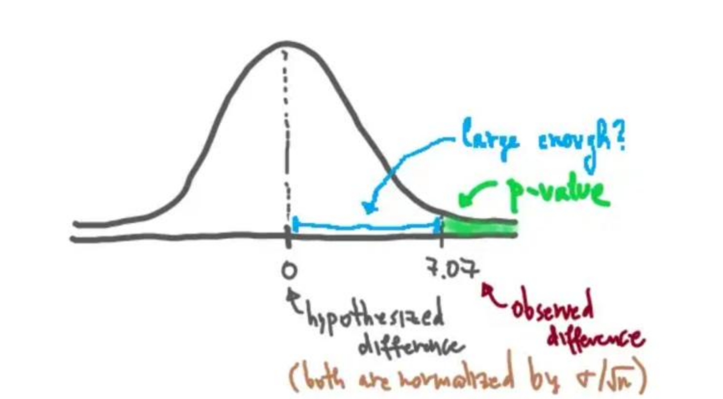
  - Part3) **rarer**

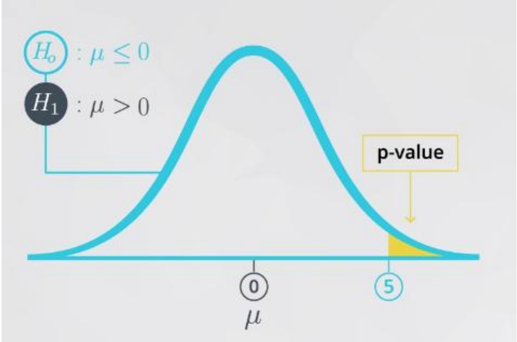
no other outcome is rarer so this is 0

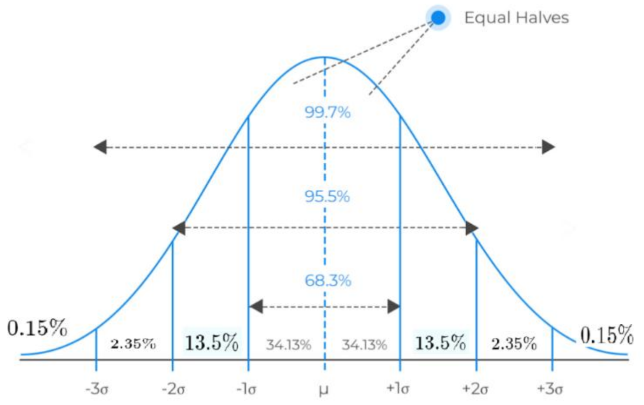
In sum, p value = 0.25 + 0.25 + 0

- Basically event (our data point being evaluated) is the line for 7.07 in below chart and green area is the “pvalue” that we need.

- pvalue is the area under the curve (/distribution) of probability of bunch of rare events on one extreme of curve. less area – thus more rare or further from mean implies that data point in question is different and not part of the population in question thus null hyp. rejected and alternate hypo accepted.
- measure of how rare an event is. more rarer is good if one is looking for rarity or difference itself from the mean
- It is also a measure of how far from the mean (mean is normal) far is weird and new/unwanted/unknown.
- 5% alpha is a threshold above which it is not that rare and thus could be within 2SDs and thus we might better as well go with the mean.
  Fun fact: according to the 3-σ rule, we can expect $\bar{\mathbf{X}}$ to be within 2 standard deviations of the expected average μ, 95% of the time.







"Measure of how far from the mean"	p-value reflects how incompatible the observed data is with H0, often linked to "distance" from the null hyp's expected mean.
"Low p-value means 'rare' and thus accept alternative hypothesis"	Correct but be cautious: ~~"Accept" H1~~ implies certainty. It's more accurate to say **"support H1"** because **statistical tests don't prove hypotheses.**
"p-value is not the 'probability' of the statistic"	True but better said as... It's the probability of observing the statistic (or more extreme values) under H0.
"statistically significant"	universally means that the p-value is less than a chosen threshold, often 0.05. It comes reused sentence that low p value means the sample data point and population mean are statistically far apart or statistically significant difference
"unlikely to have occurred by random chance" or "due to chance"	"Chance" here refers to the randomness inherent in the null model or assumption, not randomness in an absolute sense. Due to/Unlikely to chance – is referring to that the observed data or result is consistent or inconsistent with what we'd expect from random variation or random occurrence of data when collecting data. "due to chance" is said for p>0.05, that data is as expected and data point close to the mean thus H1 is negated. "unlikely random" is said for p<0.05, that data is not expected and further away from mean, thus H0 is negated. H1 is possible.

Z test, student test, KPSS test… blah blah test are all slightly different distributions (shape of the curve is different) such that value of or area changes (green area above changes)

- pvalue is supportive evidence against null hypothesis. Supportive cos high pvalue leads to accepting null hypo and low value leads to rejecting null hypothesis.
- small p value also means that old is good so it is tricky to think when as experimenter you would be rooting for rejecting null hypothesis thus a low pvalue.

why is pvalue theory so roundabout and difficult to understand?

Basically, we usually know what is known and less of the unknown so how does one estimate the unknown and tell how rare is an event?

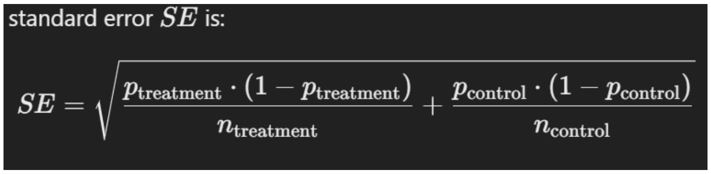
- if you won 100$ today? how lucky is that? do many people get lucky etc?

Here we assume the curve and that events distribution are as per the curve. based on the curve, the rare event (such as winning 100$) and the mean (found out by talking to atleast 30 people or known datapoints) you can calculate the pvalue and see if its outside the 2SDs. if outside 2SDs then it is rare else it is told it is common and could be due to random chance.

Numeric demo for p value:

Lets see how p value behaves in experiments.
Lets take an experiment where new drug is tested among ten people in Treatment group and 10 people given placebo in Control group.

Formulas used for below table:
1) Standard Error:

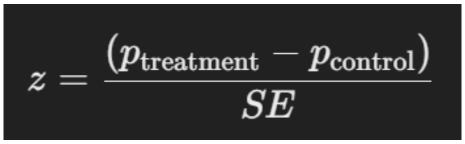


standard error SE is:

$$SE = \sqrt{\frac{p_{\text{treatment}} \cdot (1 - p_{\text{treatment}})}{n_{\text{treatment}}} + \frac{p_{\text{control}} \cdot (1 - p_{\text{control}})}{n_{\text{control}}}}$$

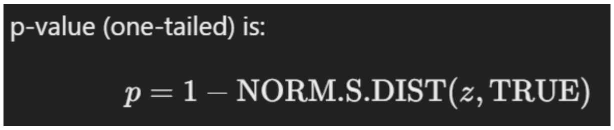
2) Test Used: A two-proportion z-test is appropriate here because we are comparing proportions between two independent groups.

3) Z Statistic



$$z = \frac{(p_{\text{treatment}} - p_{\text{control}})}{SE}$$

4) p value (Excel Formula else is looked up p value chart):



p-value (one-tailed) is:

$$p = 1 - \text{NORM.S.DIST}(z, \text{TRUE})$$

NORM.S.DIST(z,cumulative) : Returns the standard normal distribution (has a mean of zero and a standard deviation of one). Use this function in place of a manual table of standard normal curve areas.

	Experiment 1	**2**	**3**	**4**	**5**	**6**	**7**
Treatment/H1 Grp (Cured)	**8**	**8**	**6**	**2**	**3**	**4**	**4**
Treatment/H1 (Total)	10	10	10	10	10	10	10
pTreatment Proportion	=8/10=0.8	0.8	0.6	0.2	0.3	0.4	0.4
Control/H0 (Cured)	**3**	**5**	**3**	**0**	**0**	**4**	**5**
Control/H0 (Total)	10	10	10	10	10	10	10
pControl Proportion	=3/10=0.3	0.5	0.3	0	0	0.4	0.5
Standard Error	0.19	0.2	0.21	0.13	0.14	0.21	0.22
Z Statistic	2.6	1.48	1.41	1.58	2.07	0	0.45
p value	0.0047	0.069	0.078	0.057	0.019	0.5	0.32
	Significant	Not Signi	Not Signi		Significant	Not Signi	Not Signi

Inference:

- **IMPORTANT**: pvalue is a quick measure between two situations, that indicate, For eg. if we are testing a new drug then its all about if the drug cure or **not enough**. By enough, we are checking comparative performance of the 2 situations - Treatment and Control. As in, if Control had 0 or appalling results then even a small success like 2/10 could pass (as seen above) but if Control group had a few successes then the Treatment group is expected to perform higher such that the difference between Control group and Treatment is even bigger. Eg: See that when Control is 0 cured above even 2 cured for Treatment was enough but when Control cured cases go up, 2 difference is not enough but a higher difference for p value to be significant.
  p value seems to be inversely a measure of the difference in the results between Treatment and Control situations.

- **p value is significant (below 0.05) :**

  1. if the Treatment grp value is Higher/More positive than Control grp value
     - As in Exp. 6 and 7 where the final pvalues are HUGE.
     - Even if the values are equal as in Exp. 6 the p value is huge
  2. If there is sufficient gap between the higher $p_{\textbf{treatment}}$ and $p_{\textbf{control}}$
     - As seen in 2 and 3, if the gap reduces like from 1 to 2 the p value becomes big

- **p value is significant even when the real world case is a failure**

  - In 4, above two rules are followed but from a real world situation, the drug is not that viable for production if only 2 out of 10 were cured. Regardless, if it atleast cured 2 people when no other drug worked then yes it is success!.
    Hence p value being significant must be reviewed with the actual real world situation before a decision is made.

p value Interpretation Guidelines:

- A low and significant p value tells that Treatment or H1 case is better than H0 case and that the difference between the data values are big enough to give considerations to H1

- A low and significant p value means that the two cases Control H0 and Treatment H1 are different where the mean of the final results are apart from each other and that H1 values are better than the H0 values. That's all.

- Real world application: A low and significant p value is a start in right direction but does not mean that H1 is approved. The data could be like in Exp. 4 above where the overall performance is still not viable. It only suggest deeper dive into the data to see how the experiment was setup.

Statistical Power

- **Stat Power is another important metric of Experiment Design that is a quick reference.**
- **It is actually a probability percentage that is calculated for the experiment and if above 80% it is said to be really powerful/good which helps if p value is on the wall like 0.049 or 0.05 itself.**

  What about situations when H1 is True but rejected it and went with H0 since pvalue was big enough? Statistical Power is a fail safe that can be used to accept the H1 case even if p value wasn't significant.

Why Is Statistical Power Important?

A high statistical power comes to rescue for cases when p value is high but H1 case is good. It helps prevent False Negative or Type II error in Hypothesis Testing.

For example:

- If a drug actually works (H1 is true), but the test concludes it doesn't, that's a Type II error.
- High power in such a situation would lead to correct application. High power means you're more likely to correctly conclude the drug works.

How and when does power increase or ends up high:

1. Large Sample Size: Large sample sizes could show how effective H1 case is actually and allows us to overrule the pvalue
2. Low variance in the data input leads to higher power
3. One Tail test (testing for effect in one direction) has more power than Two-Tailed test
4. Increasing Significance Level: Sometimes 0.01 is used which leads to prematurely rejecting H1. Increasing it to 0.05 creates space for accepting H1
5. Large Effect Size like large number of cured patients from new drug (H1) leads to higher power. Kind of obvious but just so.
6. Overlap: Higher power happens when less overlap between two populations

How Is Statistical Power Calculated?

Statistical power is expressed as:

$$\text{Power} = 1 - \beta$$

- **β:** Probability of a Type II error (failing to reject H0 when H1 is true).
- Power is typically considered adequate if it's at least **0.8** (80%).

Example: Drug Testing

- H0: The drug has no effect.
- H1: The drug has an effect.

If: Sample size is 50, Effect size (difference in recovery rates) is large, Variance is low, α is 0.05, ...the power might be 0.9, meaning a 90% chance of detecting the drug's effect if it truly works.

To increase power:

1. **Increase sample size.**
2. **Reduce variability** in measurements.
3. **Increase effect size**, if possible (e.g., stronger treatment doses).
4. **Use a higher significance level (α),** though this increases the risk of Type I error.
5. **Use a one-tailed test** if the direction of the effect is clear.

Sample Size

Sample size is **crucial** in inferential statistics as it determines **the choice of distribution and the reliability of results**.

The number **30** is commonly cited due to the **Central Limit Theorem (CLT)**, which states that for **n ≥ 30**, the sampling distribution of the mean tends to be **approximately normal**, even if the original population is not normally distributed. However, 30 is not an absolute rule, and context matters:

What is "large" “med” "small"?

- **n < 30 (Small Sample)**
  - The sample may **not approximate normality** (unless population is already normal).
  - **T-Tests** (instead of Z-Tests) are used since the population standard deviation is usually unknown.
  - Bootstrap methods or non-parametric tests (e.g., Mann-Whitney U) may be preferred.
- **n ≥ 30 (Moderate Sample)**
  - CLT starts to apply, meaning **sample mean approximates normality** even for non-normal populations.
  - Parametric tests like Z-Test or T-Test work well.
- **n > 100 (Large Sample)**
  - The **law of large numbers** ensures more stable estimates of population parameters.
  - Normal approximations become even more reliable.
  - Large datasets allow for complex models (e.g., multiple regression, machine learning).

So, is 30 always enough?

- For normal or near-normal populations → Yes, 30+ is reasonable.
- For skewed or heavy-tailed distributions → You may need n > 50 or even 100+.
- For rare events (e.g., employee attrition in a small company) → Larger samples (n > 100) are often needed.

Connecting the Dots in Inferential Statistics

The Big Picture

Inferential statistics exists to solve a fundamental problem: we usually can't measure an entire population, but we need to have universal framework (so that others can also verify a claim or redo the test) to make reliable conclusions about that population. This requires us to:

1. Ensure we have population stats known already
2. Take a current sample from the population possibly after a treatment
3. Measure properties of that sample
4. Make inferences about the wider population
5. Quantify our confidence in those inferences

Each concept , **Probability Distribution, CLT, Sample size , p value and hypothesis testing** play a critical role in this process.

How These Concepts Connect

TLDR:

1. LLN says that if sample size keeps increasing the mean of sample is very close or say if sample size is largest, it is the population mean.
2. CLT says that means of any new sample (as long it is big enough and unbiased) including above big sample are together in one normal distribution. Empirically this is proven.
3. Because they are together and we know its a normal curve, we can see if they are together or far apart by finding p value.
4. Together means they are similar and like in same family /nothing has changed and if far apart they are like two different families and certainly not similar

Probability Distributions: The Mathematical Models of Variation

Probability distributions are mathematical models that describe how values in a population are distributed. The normal distribution is particularly important because:

- It occurs naturally in many real-world phenomena (heights, measurement errors, etc.)
- It has mathematical properties (eqn can be defined by an eqn and thus we can use this to calculate area of a part of the curve which is the probability) that make analysis simpler
- Most importantly: even when the underlying population isn't normally distributed, the **sampling distribution** of many statistics (like means) will be approximately normal under certain conditions

This last point connects directly to the Central Limit Theorem.

The Central Limit and Law of Large Numbers Theorems: The Bridge Between Samples and Populations

The Central Limit Theorem (CLT) states that when you take multiple samples from any population and calculate the mean of each sample, these different means , once charted shows a normal distribution, regardless of the population's original distribution.

Law of Large Numbers (LLN) tells us that as sample size increases, the sample mean will almost surely converge to the population mean. Thus any measurement taken of a large sample size should be very close to the Population mean, in fact we can consider it as population mean.

Now, in a real world situation, we probably have many such population statistics. Example: the average productivity of all people in a company to be 50 tasks per week and std of 5 tasks. Now, in comes a training or technology that is supposed to make them more productive. We test it with a small team before buying the tech & training the staff. After the training and tech deployment, we want to find if things changed. This is where CLT and LLN comes in to bridge.

First, LLN tells that a mean of a super large sample should be super close to population mean. Thus we can even consider that to be the population mean.

Second, CLT says all the means of samples end up in a normal distribution, thus the population mean above and new mean from new sample are in same normal distribution.

This is how LLN and CLT theorems bridge sample and population together.

Since population mean and sample mean are under the same normal distribution, we can then check if they are close to each other or far apart by seeing how many stds are between them which help us conclude that they are in the same distribution or not in the same distribution but different distributions. More on this later.

Coming back to our example, all we have are the “weekly” population parameters. So lets take samples by measuring “weekly” for 4 weeks, the productivity of the test team. This will give us the sample mean. Using the population mean and pop. Std, we do standardization and calculate the z score. The z score is from the z table which is nothing but probability values found by measuring area under the curve. The z score is the p value which is probability that indicates if the sample value is close to population mean or far apart.

CLT/LLN are incredibly powerful that it allows us:

- It lets us use normal distribution mathematics even when our original data isn't normal
- It provides a mathematical foundation for how sample statistics relate to population parameters
- It explains why larger samples give more reliable estimates

Think of it as a mathematical miracle that allows us to make reliable inferences despite having incomplete information.

Sample Size and Unbiased Sampling: The Foundation of Validity

For the CLT to work properly, we need:

1. **Sufficiently large samples**: Larger samples better approximate the population and produce sampling distributions that are more normal

2. **Unbiased sampling**: Every member of the population must have an equal chance of being selected

Without these conditions, our statistical inferences could be systematically wrong. This is why proper sampling techniques are emphasized so heavily in statistics courses.

Hypothesis Testing and p-values: The Decision-Making Framework

Once we have our sample and understand how it relates to the population (through probability distributions and the CLT), we need a framework for making decisions. This is where hypothesis testing comes in:

1. We state a null hypothesis (typically that "nothing interesting is happening")
2. We collect sample data
3. We calculate how likely our sample result would be if the null hypothesis were true
4. This likelihood is quantified as the p-value

The p-value represents the probability of getting a result at least as extreme as our sample, assuming the null hypothesis is true. If this probability is very small (typically < 0.05), we reject the null hypothesis.

R2

(Refer to ML Modelling Notes for more)
Tells the amount of variance explained by the best fitting regression line against one a line that is just mean of the data. It tells how much error was reduced due to the regression line thus higher R2 means higher error reduction and closer fit to data

Adjusted R squared

(Refer to ML Modelling Notes for more)
Adjusted R2 is an attempt to take account of the phenomenon of the R2 automatically and spuriously increasing when extra explanatory variables are added to the model.

Test Types:

Type	Examples	Purpose
One-Sample Tests	z-test (if n>30), t-test (if n<30)	Compare sample to a known value
Two-Sample Tests	Independent t-test, Paired t-test	Compare two groups

Multi-Sample Tests	ANOVA, Kruskal-Wallis Test	Compare three or more groups
Tests for Relationships	Correlation, Regression, Chi-Square	Test for relationships or dependencies
Tests for Distribution	Shapiro-Wilk, KS Test	Test if data follow a distribution
Tests for Variance	Levene's Test, F-Test	Test if variances are equal
Non-Parametric Tests	Mann-Whitney, Wilcoxon	Test when data do not meet parametric assumptions
Tests for Proportions	Z-Test for Proportions, Chi-Square	Compare proportions in groups

1. One-Sample Tests

- **Definition**: Compare a single sample to a known population value or theoretical benchmark.
- **Use Case**: When you have one group and want to test it against a fixed standard.
- **Examples**:
  - **One-Sample t-Test**: Test if the sample mean differs from a known population mean.
  - **One-Sample z-Test**: Test for a sample mean when the population standard deviation is known.
  - **One-Sample Proportion Test**: Test if a sample proportion differs from a hypothesized proportion.

2. Two-Sample Tests

- **Definition**: Compare two independent or dependent samples to test for differences in their means, variances, or proportions.
- **Use Case**: Comparing two groups to determine if they are statistically different.
- **Examples**:
  - **Independent Two-Sample t-Test**: Compare means of two independent groups.
  - **Paired t-Test (Dependent t-Test)**: Compare means of two related groups (e.g., before and after a treatment).
  - **Two-Sample Proportion Test**: Compare proportions between two groups.
  - **F-Test for Equality of Variances**: Compare variances between two groups.

3. Multi-Sample Tests

- **Definition**: Compare more than two groups to test for differences.
- **Use Case**: Analyzing more than two groups simultaneously to detect differences.

- **Examples**:
    - **ANOVA (Analysis of Variance)**: Tests if the means of three or more groups are significantly different.
        - **One-Way ANOVA**: Tests one factor (e.g., treatment type) across multiple groups.
        - **Two-Way ANOVA**: Tests two factors (e.g., treatment and time) across groups.
    - **Kruskal-Wallis Test**: Non-parametric test for differences among three or more groups.

4. Tests for Relationships

- **Definition**: Assess relationships or associations between variables rather than differences.
- **Use Case**: When interested in how variables are related or dependent on each other.
- **Examples**:
    - **Chi-Square Test of Independence**: Tests if two categorical variables are independent.
    - **Correlation Test (Pearson/Spearman)**: Tests the strength and direction of the relationship between two variables.
    - **Regression Analysis**: Tests how one or more independent variables predict a dependent variable.

5. Tests for Distribution Fit

- **Definition**: Check if a sample matches a theoretical distribution.

- **Use Case**: Testing assumptions about the data distribution.
- **Examples**:
    - **Kolmogorov-Smirnov Test**: Tests if a sample follows a specific distribution (e.g., normal distribution).
    - **Shapiro-Wilk Test**: Tests if data are normally distributed.

6. Tests for Variance

- **Definition**: Compare variability (spread) within or between groups.
- **Use Case**: When variance is of interest rather than the mean.
- **Examples**:
    - **Levene's Test**: Tests for equal variances between groups.
    - **Bartlett's Test**: Tests for equal variances (for normally distributed data).

7. Non-Parametric Tests

- **Definition**: Tests that do not assume data follow a specific distribution (e.g., normal distribution).
- **Use Case**: When data violate parametric test assumptions or are ordinal.
- **Examples**:
    - **Wilcoxon Signed-Rank Test**: Non-parametric alternative to the paired t-test.
    - **Mann-Whitney U Test**: Non-parametric alternative to the independent t-test.

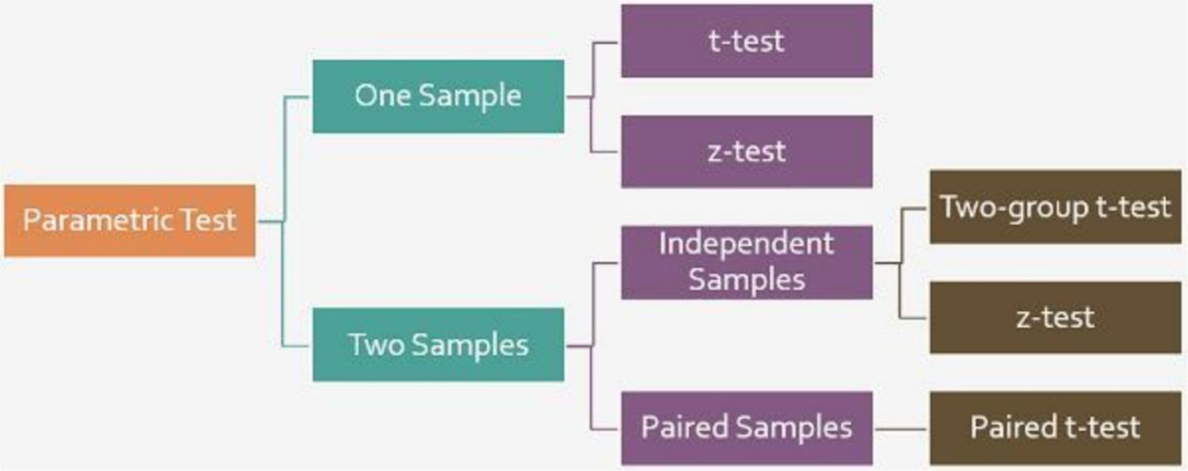
    - **Kruskal-Wallis Test**: Non-parametric alternative to one-way ANOVA.

8. Tests for Proportions

- **Definition**: Compare proportions within or across groups.
- **Use Case**: When data are in categorical form.
- **Examples**:
    - **Chi-Square Goodness-of-Fit Test**: Tests if observed proportions match expected proportions.
    - **Z-Test for Proportions**: Compares proportions between two groups.

Application	Test Names	Statistical Test Type
Compare hiring success rates before and after implementing a new recruitment strategy	Z-Test / T-Test	One Sample Test
Compare hiring pass rates across demographic groups	Chi-Square Test	Test for Independence
Measure employee performance before and after training	Paired T-Test	Dependent Sample Test
Compare multiple training methods to find the most effective one	ANOVA	Variance Comparison
Compare attrition rates across different departments or demographic groups	Chi-Square Test	Test for Independence
Compare engagement scores between different teams, gender groups, or remote vs. in-office employees	T-Test / ANOVA	Mean Comparison Test
Analyze survey response distributions (e.g., % satisfied vs. dissatisfied across locations)	Chi-Square Test	Test for Independence
Compare productivity before vs. after implementing a new policy (e.g., remote work)	T-Test	Dependent Sample Test

Major Dichotomies in Test Types:



0: One Sample vs Two Sample vs Multisample as above

1. **Paired (Dependent) vs. Independent Samples**

- **Independent Samples**: When the samples are completely unrelated to each other (e.g., heights of men vs. women).
- **Paired (Dependent) Samples**: When the samples are related, such as before-and-after measurements or matched pairs.

2. **One-Sided vs. Two-Sided Tests**

- **One-Sided Test**: Tests whether a sample statistic is greater than or less than a population parameter or another sample statistic in one specific direction.
  - Example: Testing if a new drug improves recovery times (but not if it worsens them).
- **Two-Sided Test**: Tests whether a sample statistic is either greater or less than a population parameter or another sample statistic, without specifying the direction.
  - Example: Testing if a new drug has any effect (better or worse).

3. **Parametric vs. Non-Parametric Tests**

- **Parametric Tests**: Assume data follows a specific distribution (often normal). Examples:
  - z-test
  - t-test
  - ANOVA
- **Non-Parametric Tests**: Do not assume a specific distribution. Useful for ordinal data or when assumptions of parametric tests are violated. Examples:
  - Mann-Whitney U Test
  - Wilcoxon Signed-Rank Test
  - Kruskal-Wallis Test
  - Chi-Square Test

z Test / T test/ Chi Sq Test/ ANOVA

1. Z-Test

- **When to Use**:
  - Large sample sizes (n>30).
  - Population variance is known.
  - Testing means for one or two samples.
- **Distribution: Standard Normal (Z-distribution, mean = 0, std = 1).**
- **Example**: Testing if the average weight of a population differs from a known value.
- **Interview Q&A**:
  - *Q: When do you use a Z-test?*
    A: When the sample size is large, and the population standard deviation is known.
  - *Q: What is the significance of the Z-distribution?*
    A: It standardizes the data, allowing comparison across different datasets.

2. T-Test

- **When to Use**:
  - Small sample sizes (n<30).
  - Population variance is unknown.
  - Comparing means between one, two independent groups, or paired samples.
- **Distribution: Student's t-distribution (similar to normal but with heavier tails for small n).**
- **Types**:
  - **One-sample t-test**: Compare sample mean to a known mean.
  - **Independent two-sample t-test**: Compare means of two independent groups.
  - **Paired t-test**: Compare means of related groups (e.g., before-and-after tests).
- **Example**: Testing if the test scores of two teaching methods differ.
- **Interview Q&A**:
  - *Q: How does a T-distribution differ from a Z-distribution?*
    A: t-distribution has heavier tails, accounting for small sample sizes.
  - *Q: What assumptions are needed for a t-test?*
    A: Normality of data, independence of observations, and equal variances (for two-sample).

3. ANOVA (Analysis of Variance)

- **When to Use**:

  - Comparing means of three or more groups.
  - Testing if at least one group mean is different.
- **Distribution: F-distribution.**
- **Types**:
  - **One-way ANOVA**: Single factor, multiple groups.
  - **Two-way ANOVA**: Two factors, with or without interaction.
- **Example**: Testing if three fertilizers result in different crop yields.

- **Interview Q&A**:

  - *Q: Why use ANOVA instead of multiple t-tests?*
    A: To control the overall Type I error rate.
  - *Q: What are the assumptions for ANOVA?*
    A: Independence, normality, and equal variances.

4. Chi-Square Test

- **When to Use**:
  - Testing relationships between categorical variables.
  - Goodness-of-fit test: Does observed data fit expected distribution?
  - Test for independence: Are two categorical variables independent?
- **Distribution**: Chi-square distribution.

- Example: Testing if gender and voting preference are independent.

- **Interview Q&A**:

- *Q: What are the conditions for a Chi-square test?*
  A: Sufficiently large sample, expected frequencies > 5.
- *Q: Can Chi-square test be used for numerical data?*
  A: No, it is for categorical data.

5. F-Test

- **When to Use**:
  - Comparing variances between two groups.
  - Used in ANOVA and regression analysis.
- **Distribution**: F-distribution.

- **Example**: Testing if two machines have similar variability in production.

- **Interview Q&A**:

- *Q: What does the F-statistic measure?*
  A: Ratio of variances between groups.
- *Q: What are the assumptions for the F-test?*
  A: Normality of data and independence of observations

6. Kolmogorov-Smirnov Test

- **When to Use**:
  - Comparing a sample distribution to a reference distribution.
  - Two-sample version tests if two distributions are the same.
- **Example**: Testing if sales follow a normal distribution.
- **Interview Q&A**:
  - *Q: What does the KS test check for?*
    A: Differences in distributions.

Test	Purpose	Key Assumptions	Distribution
Z-Test	Test mean for large n	Normality, known variance	Normal
T-Test	Test mean for small n	Normality, unknown variance	Student's t
ANOVA	Compare means of 3+ groups	Normality, equal variances	F-distribution
Chi-Square	Categorical variable relationships	Expected f>5, independence	Chi-square
F-Test	Compare variances	Normality	F-distribution
Binomial	Proportion for binary outcome	Binary data	Binomial
Wilcoxon/Mann-U	Non-parametric mean/median comparison	No normality required	Rank-based
Kruskal-Wallis	Non-parametric ANOVA alternative	No normality required	Rank-based

6. Binomial Test

- **When to Use**:

    - Testing proportions in a binary outcome.
- **Distribution**: Binomial distribution.
- **Example**: Testing if a coin is fair.
- **Interview Q&A**:
    - *Q: When do you use a binomial test?*
      A: When testing a proportion against a fixed value.

7. Wilcoxon and Mann-Whitney U Tests

- **When to Use**:
    - Non-parametric alternatives to t-tests.
    - Data is not normally distributed.
- **Distribution**: Based on ranks, not data values.
- **Example**: Comparing median scores of two groups.
- **Interview Q&A**:
    - *Q: When would you use a Wilcoxon test?*
      A: When the data violates the t-test assumptions.

8. Kruskal-Wallis Test

- **When to Use**:
    - Non-parametric alternative to ANOVA.
    - Comparing medians of three or more groups.
- **Distribution**: Based on ranks.
- **Example**: Testing if medians of three teaching methods differ.
- **Interview Q&A**:
    - *Q: Why use Kruskal-Wallis instead of ANOVA?*
      A: When data is not normally distributed.

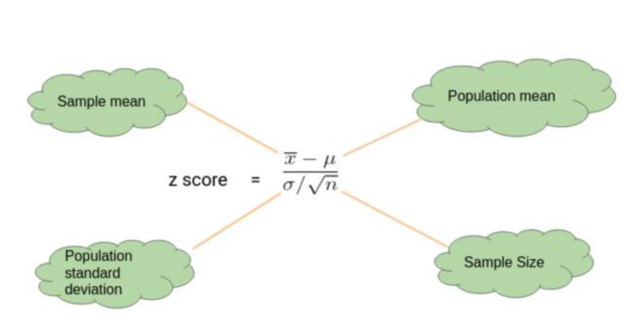
https://medium.com/@jw207427/how-to-apply-hypothesis-test-in-marketing-data-fbe1e1ac2388 or in online resources page

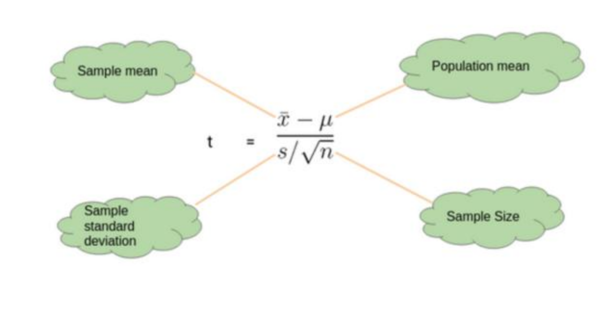
z test	**t test – small (small as in less than 30 sample)**	**chi sq (correlation between two datasets)**	**ANOVA**	**f test**
Testing means for one or two samples.	Comparing means between one, two independent groups, or paired samples.			
when sample size is large (n>30), or the population	sample size is small (n<30) and population SD is	two data sets/ two sample proportion/two	comparison of three or more means/samples	two variances

SD (or var as mentioned in books) is known	**not known**	categorical		
Eg: Testing if the average weight of a population differs from a known value.				
abt Population	abt Sample		abt Sample variances	abt one sample variance
	Best for two samples (old sample and new markt campaign sample)			
	If the two samples are dependent then do paired t test if independent then normal t test			
	one sample proportion usually one categorical like gender (M:F)	two **proportions** like m:f ratio for age are dependent. male/female prop is dependent on age. That is m:f is different for kids, adults and oldies. thus two props relation is tested	when a categorical has more than 2 groupings like age (young, adult and old). if two like m:f you do t-test	
	one numeric		Both compare variances	

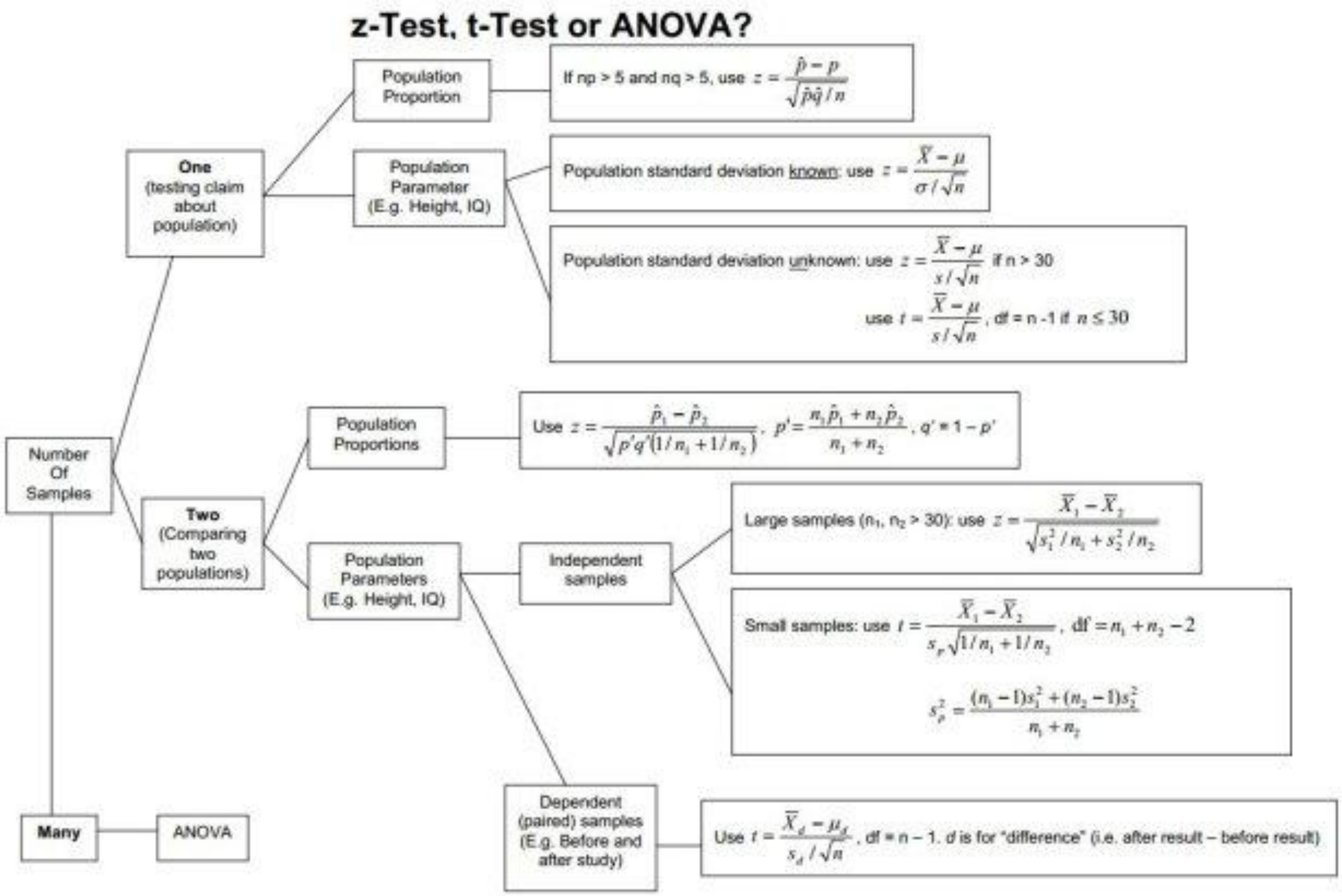
Theory:

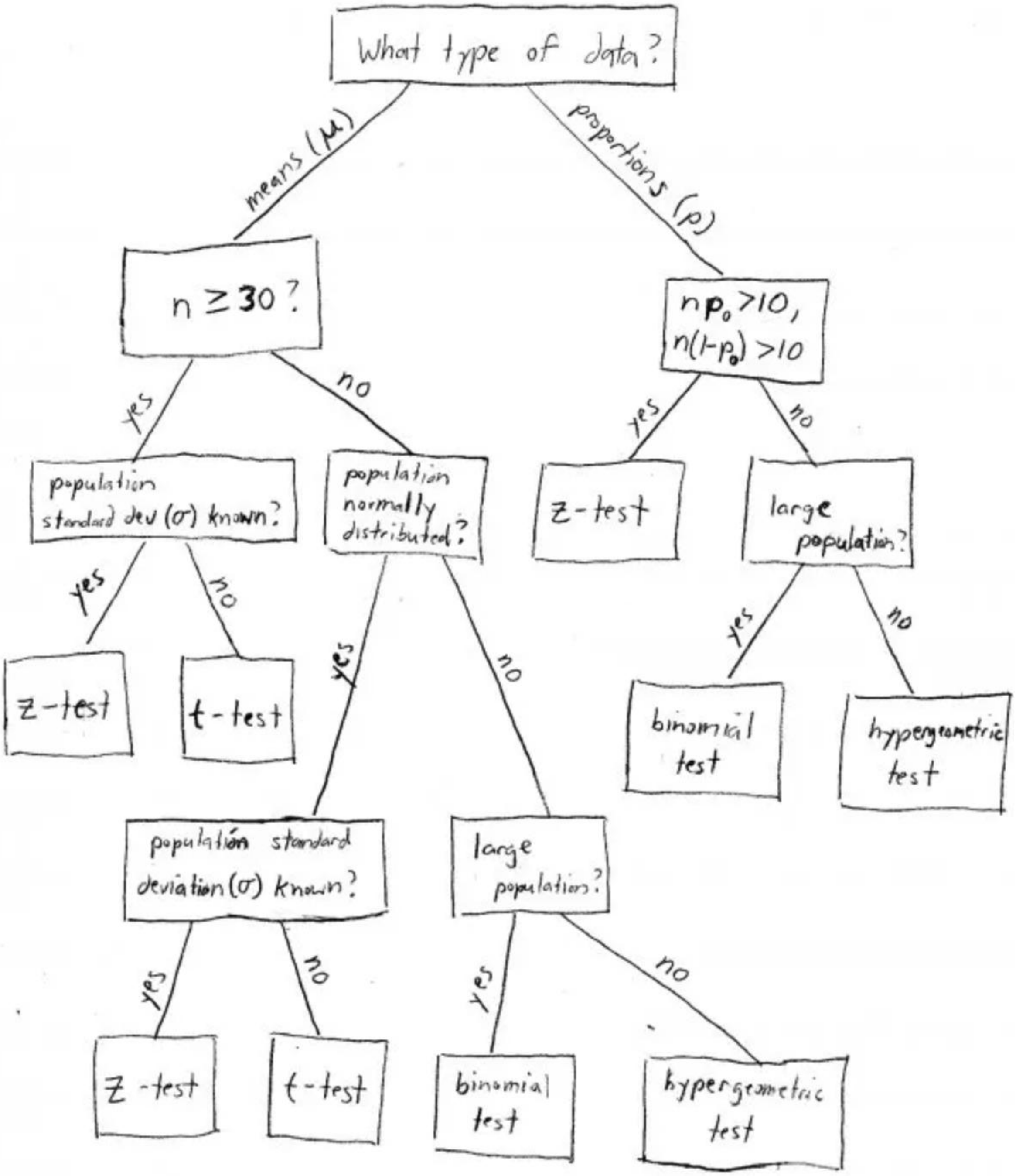
Formula:


Sample mean
Population mean
z score = $\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$
Population standard deviation
Sample Size


Sample mean
Population mean
t = $\frac{\bar{x}-\mu}{s/\sqrt{n}}$
Sample standard deviation
Sample Size

$z = \frac{observed - expected}{SE}$	
Formula for z and t score are ALMOST same. - Z uses population SD while t uses sample SD - difference is in the normal distribution shape with student t distribution is shorter in middle and fatter in tails.	
Chi- square: $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ where O_i = observed value (actual value) E_i = expected value	$\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

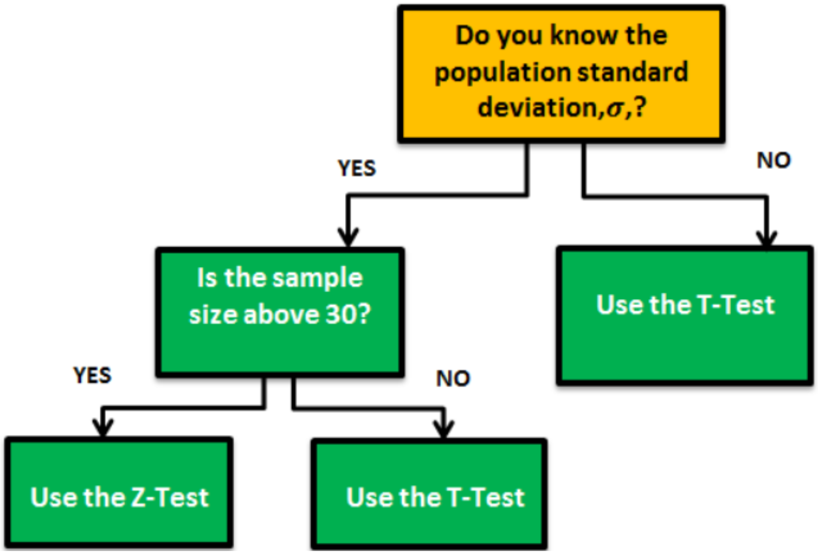

z-Test, t-Test or ANOVA?
Number Of Samples
One (testing claim about population)
Population Proportion
If np > 5 and nq > 5, use $z = \frac{\hat{p} - p}{\sqrt{\hat{p}\hat{q}/n}}$
Population Parameter (E.g. Height, IQ)
Population standard deviation known: use $z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$
Population standard deviation unknown: use $z = \frac{\overline{X} - \mu}{s/\sqrt{n}}$ if n > 30
use $t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$, df = n -1 if $n \leq 30$
Two (Comparing two populations)
Population Proportions
Use $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p'q'(1/n_1 + 1/n_2)}}$, $p' = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$, $q' = 1 - p'$
Population Parameters (E.g. Height, IQ)
Independent samples
Large samples ($n_1, n_2 > 30$): use $z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$
Small samples: use $t = \frac{\overline{X}_1 - \overline{X}_2}{s_p\sqrt{1/n_1 + 1/n_2}}$, $\text{df} = n_1 + n_2 - 2$
$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}$
Dependent (paired) samples (E.g. Before and after study)
Use $t = \frac{\overline{X}_d - \mu_d}{s_d/\sqrt{n}}$, df = n – 1. d is for "difference" (i.e. after result – before result)
Many
ANOVA

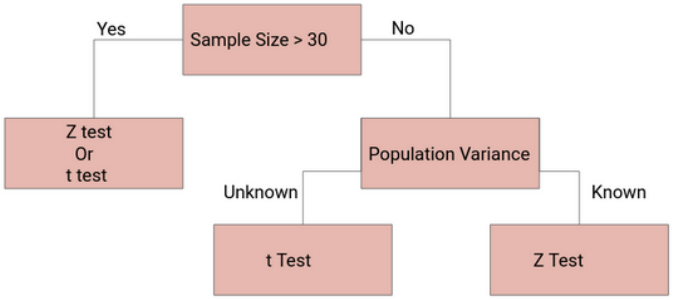

What type of data?
means (μ)
proportions (p)
n ≥ 30?
yes
no
population standard dev (σ) known?
yes
no
z-test
t-test
population normally distributed?
yes
no
population standard deviation (σ) known?
yes
no
Z-test
t-test
large population?
yes
no
binomial test
hypergeometric test
np₀ > 10, n(1-p₀) > 10
yes
no
z-test
large population?
yes
no
binomial test
hypergeometric test

Deciding between Z Test and T-Test

normal (Gaussian) distribution curve does a good job in approximating the distribution of a sample statistic (such as the mean). It does not, however, do a good job in approximating the distribution of that same statistic when these are computed from small sample sizes.

So when we should perform the Z test and when we should perform t-Test? It's a key question we need to answer if we want to master statistics.

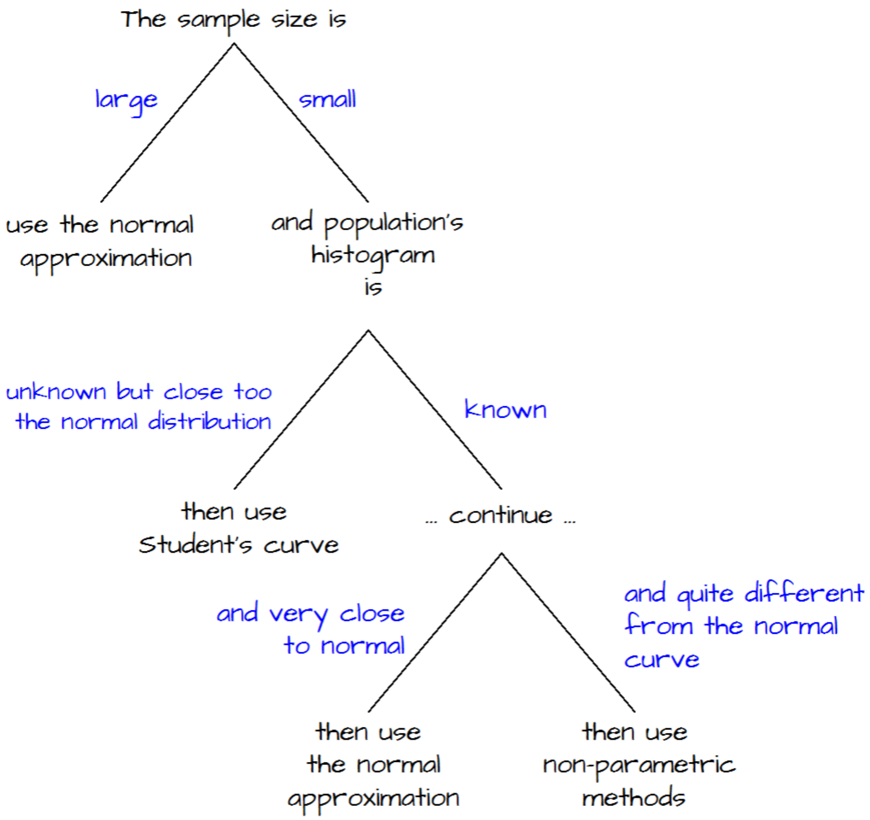

Do you know the population standard deviation,σ,?
YES
NO
Is the sample size above 30?
Use the T-Test
YES
NO
Use the Z-Test
Use the T-Test



If the sample size is large enough, then the Z test and t-Test will conclude with the same results. For a **large sample size**, **Sample Variance will be a better estimate** of Population variance so even if population variance is unknown, we can **use the Z test using sample variance.**

Similarly, for a **Large Sample**, we have a high degree of freedom. And since t**-distribution approaches the normal distribution**, the difference between the z score and t score is negligible.

Which curve to use:


The sample size is
large
small
use the normal approximation
and population's histogram is
unknown but close too the normal distribution
known
then use Student's curve
... continue ...
and very close to normal
and quite different from the normal curve
then use the normal approximation
then use non-parametric methods

Null hypothesis of different tests:

Pearson's correlation test :

the null hypothesis for the Pearson's correlation test is that there is no relationship between two variables.

Student's t test:

The null hypothesis for the Student's t test is that there is no difference between the means of two populations.

Chi-Square Goodness of Fit Test – Used to determine whether or not a categorical variable follows a hypothesized distribution.

Chi-Square Test of Independence – Used to determine whether or not there is a significant association between two categorical variables.

Degrees of Freedom:

Intuition:

Is the number of independent data points we have on making an estimate (stat of a sample and distinguishing that it is not of population).

Eg: How many spikes does a sea urchin have? you take 5 urchins and count.
Here when you calculate sample mean, the population mean is said to have 5 DF.

Once SD of samples are done then the population SD is said to have 4DF.

In same fashion, Skewness has 3DF and Kurtosis has 2DF.

Experimental Design

ED is about careful planning of conducting the experiment and involves hypothesis formulation, bias control, statistical tests, and practical significance to ensure **valid, reliable, and interpretable** results

Types of Experimental Design in People Analytics

1. **Between-Subject Design**
   - **Definition**: Different groups with different treatment: like control and treatment grps I n A/B testing
   - **Use Case**: Testing two different onboarding programs by assigning new hires randomly into two groups and comparing performance after 3 months.
   - **Pros**: No carryover effects, straightforward.
   - **Cons**: Requires a larger sample size.
2. **Within-Subject Design**
   - **Definition**: The same subjects in different situations – before/after etc
   - **Use Case**: Measuring employee productivity before and after implementing a hybrid work policy.
   - **Pros**: Requires fewer participants, reduces variability.
   - **Cons**: Potential carryover effects (e.g., employees adjusting to new policies).
3. **Factorial Design**
   - **Definition**: Tests multiple independent variables simultaneously to see their combined effects.
   - **Use Case**: Evaluating how both training intensity (low vs. high) and mentorship (with vs. without) impact employee retention.
   - **Pros**: Detects interaction effects.
   - **Cons**: Can become complex with too many factors.
4. **Quasi-Experimental Design**

- **Definition**: Similar to experiments but lacks **random assignment** due to practical constraints.
- **Use Case**: Analyzing the impact of leadership training when assignment is based on voluntary participation.
- **Pros**: Useful when randomization is impossible.
- **Cons**: Higher risk of confounding variables.

Type of Quasi : Difference-in-Differences (DiD):

- A quasi-experimental technique used to estimate the causal effect of an intervention by comparing the change in outcomes over time between a treatment group and a control group.
- DiD requires both **pre** and **post** data from two groups: one that receives the treatment and one that does not. It is often used when random assignment isn't feasible (e.g., evaluating the effect of a new HR policy in one department compared to another).

Each design type is chosen based on feasibility, ethical considerations, and business constraints.

Key Considerations in Experiment Design

Factor	Why It Matters?
1. Clear Hypothesis & Objectives	Define a precise **null** and **alternative hypothesis** to ensure the experiment has a clear goal.
2. Sample Size & Power Analysis	Ensure enough observations to detect meaningful effects while avoiding under/overfitting. Power analysis helps determine the required sample size.
3. Randomization	Randomly assign participants to avoid bias and confounding variables.
4. Control Groups	A control group (e.g., A/B testing) allows for proper comparison and ensures the effect is due to the intervention.
5. Blinding (Single, Double, Triple)	Reduces bias in subjective measurements (e.g., self-reported employee satisfaction).
6. Confounding Variables & Covariates	Identify and control external factors that may influence results (e.g., seasonal trends, economic shifts).
7. Experimental Design Type	Choose from **between-subject, within-subject, factorial designs, or quasi-experiments** based on constraints.

Factor	Why It Matters?
8. Data Collection & Measurement Consistency	Ensure data is collected consistently and accurately to avoid biases.
9. Handling Missing Data	Decide on imputation strategies or whether to remove missing data based on the nature of the dataset.
10. Statistical Test Selection	Select **parametric (t-test, ANOVA) or non-parametric (Mann-Whitney, Kruskal-Wallis)** tests based on data distribution.
11. Multiple Testing Corrections	When running multiple hypothesis tests, control for false positives (e.g., Bonferroni correction).
12. Effect Size vs. Statistical Significance	A small p-value doesn't always mean practical importance—effect size measures real-world impact.
13. Time Horizon & Duration	Ensure the study runs long enough to capture effects (especially in ML where drift can occur).
14. ML-Specific Concerns	Check for **data leakage, feature selection bias, and generalization errors** in ML experiments.
15. Ethical Considerations	Ensure fairness, privacy, and ethical treatment of participants (especially in HR/People Analytics).

Selection Bias

What is Selection Bias?

Selection bias occurs when the sample used in a study **is not representative of the population**, leading to **biased conclusions**. This can happen due to **non-random sampling, self-selection, or missing data**.

General Types of Selection Bias:

1. **Sampling Bias** – The sample systematically differs from the population.
    - *Example:* Conducting an employee satisfaction survey but only collecting responses from managers.
2. **Self-Selection Bias** – Participants opt into the study in a non-random way.
    - *Example:* Only highly engaged employees participate in an employee engagement survey.
3. **Survivorship Bias** – The analysis only includes "successful" cases.

    - *Example:* Studying only employees who have stayed at the company for 5+ years to analyze retention factors.
4. **Attrition Bias** – Participants drop out of the study in a non-random way.
    - *Example:* Employees leaving the company don't complete the final pulse survey, skewing results toward happier employees.

Ways to Mitigate Selection Bias :

- **Use Random Sampling** – Ensure diverse representation across all levels, departments, and demographics.
- **Stratified Sampling** – Divide employees into meaningful groups (e.g., tenure, job role) before analysis.
- **Propensity Score Matching** – Compare similar employees by matching them on key attributes before evaluating differences. Do an experiment, match subjects or group them before analysing impact of results
- **Control for Confounders** – Use statistical techniques (like regression) to adjust for pre-existing differences.
- **Randomized Controlled Trials** (RCTs) - randomly assigning employees to training vs. control groups so that an experiment by nature does not pul only one crowd – survey is filled only by top engaging employees
- **Survey Weighting** - If only engaged employees complete the survey, results will be artificially positive. **Solution:** Use **survey weighting**, adjusting results based on known engagement levels from a random sample of employees.
- **Quasi-experimental designs** - If only managers interested in leadership development participate, the results may not generalize to all managers. Use **quasi-experimental designs** by comparing participants with a control group of similar managers using regression adjustment.

Causal Inference

Causal inference is the process of **determining whether a relationship between two variables is causal (i.e., one variable directly affects another) rather than just correlational**. Since real-world data is often **observational rather than experimental**, causal inference methods help control for confounders and biases to approximate a true cause-and-effect relationship.

Types of Causal Inference Methods

1. Experimental Methods (Gold Standard for Causality)

These involve **randomization**, which ensures that treatment and control groups are comparable.

- **Randomized Controlled Trials (RCTs)** – Participants are randomly assigned to treatment and control groups.

2. Quasi-Experimental Methods (Mimic RCTs with Observational Data)

Used when randomization is not possible, but natural or policy-based variations allow for causal inference.

- **Difference-in-Differences (DiD)** – Compares before-after changes between a treatment and a control group. **DiD** compares the change in outcomes between two distinct groups, treatment and a control, to isolate the treatment effect.
- **Regression Discontinuity Design (RDD)** – Exploits cutoff-based assignments (e.g., a policy applying only to people above a certain score).
- **Instrumental Variables (IV)** – Uses a third variable (instrument) that affects the treatment but is unrelated to confounders.

3. Matching Methods (Balancing Treatment & Control Groups)

Used when experimental data is unavailable; tries to make groups more comparable.

- **Propensity Score Matching (PSM)** – Matches individuals with similar characteristics between treatment and control groups.
- **Exact Matching** – Matches individuals with identical characteristics.

4. Structural Causal Models (SCM) & Graphical Methods

Used to formally define causal relationships using directed graphs.

- **Directed Acyclic Graphs (DAGs)** – Visualize causal structures and help identify confounders.