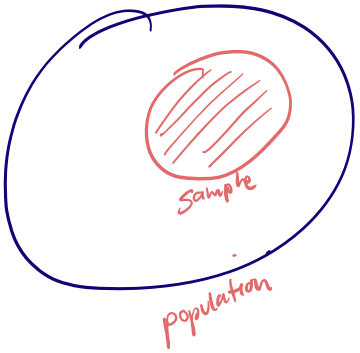


- Based on our sample we hope to get a good "estimate" for a summary of the population

- A summary about the population is a **parameter**, and a summary for our sample is a **statistic**

- How good is our statistic in estimating our parameter?

Parameter	Statistic
μ = mean	$\bar{x}$ = sample mean
p = proportion	$\hat{p}$ = sample proportion

In the case where we use a single estimate to represent the entire population, is referred to as a **point estimate**

Alternatively, we can use an Interval estimate to specify plausible values for our parameter of Interest

- There will always be an Error from our estimate. our goal is to reduce possible errors that can occur. such as

Sampling Error: Describes how much an estimate will vary from Sample to Sample. One goal is to quantify such error using the Sample size "n"

BIAS: Inappropriate Sampling Techniques, such as Convinience Sampling

* whenever Simple Random Sample is implemented observations from our Sample are considered Independent

Normal Approximation to the Binomial Distn:

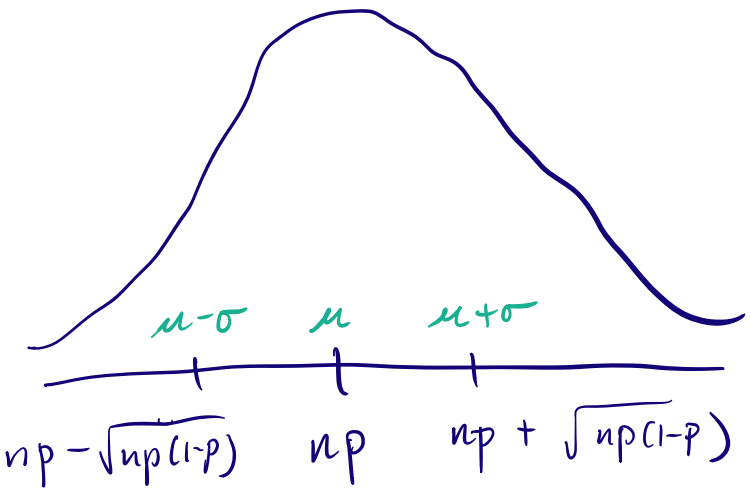
- The Binomial distribution with probability of success p is approximately normal whenever our sample size "n" is sufficiently large.

- We consider sufficiently large whenever

$$np \geq 10 \quad \text{and} \quad \underbrace{np(1-p)}_{\text{failure}} \geq 10$$

If the above conditions hold then the above Binomial distn will approximately be normal with parameters

$$\mu = np \quad \text{and} \quad \sigma = \sqrt{np(1-p)}$$



It Turns out this is only one small piece of an Important Theorem which provides a foundation for most of statistics.

Central Limit Theorem

- Recall, The Law of Large Numbers which states as we increase the number of experiments we perform the sample mean will converge to the theoretical mean.

$$\bar{X} \longrightarrow \mu \quad \text{as } n \to \infty$$

Likewise, for Bernoulli trials where we consider the proportion of successes

$$\hat{p} \longrightarrow p \quad \text{as } n \to \infty$$

where $\hat{p}$ is the proportion of successes from our sample

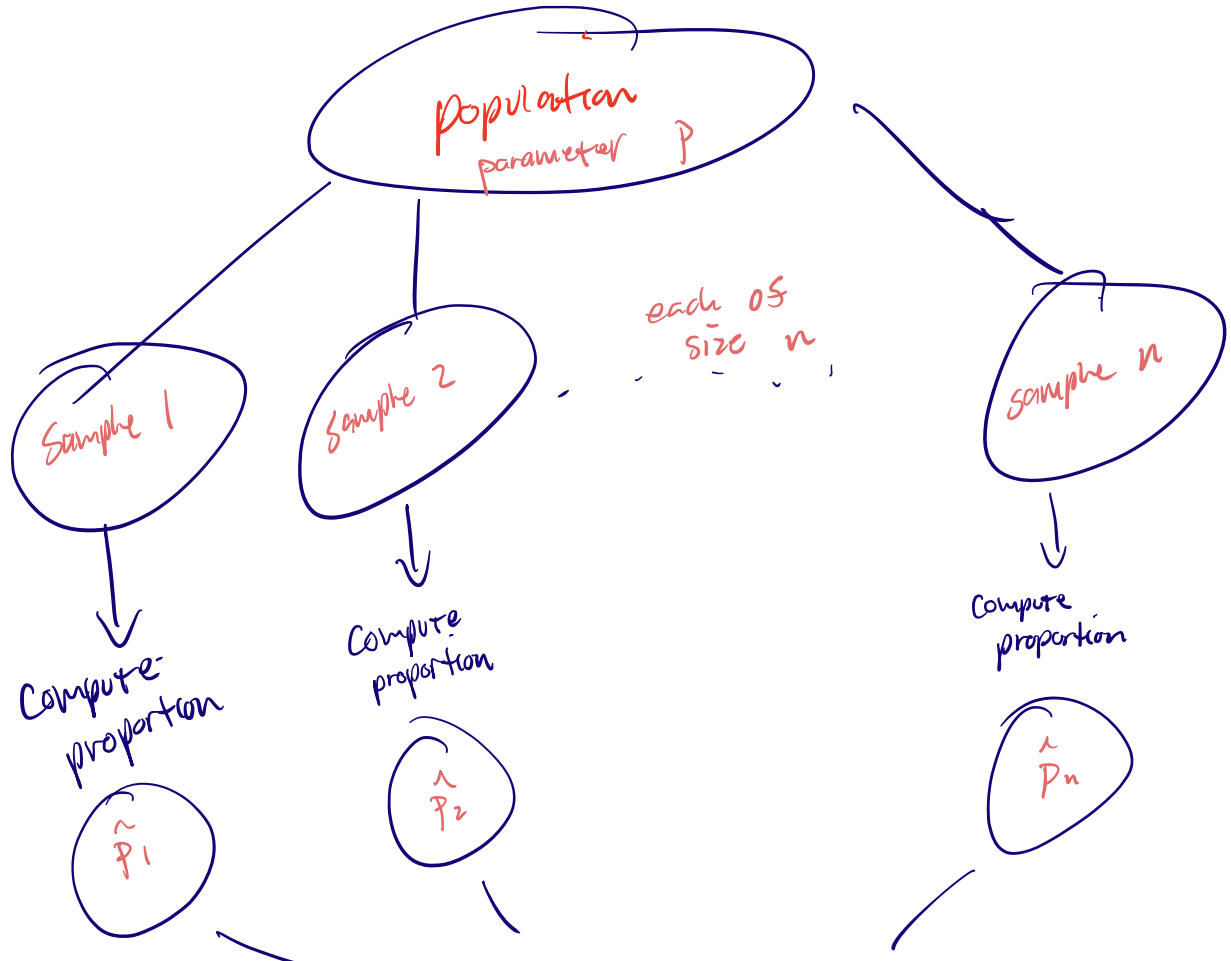
Consider proportions first, we'll do Means later.

Central Limit Theorem:

Whenever our observations are Independent and the Sample Size is sufficiently large enough ($np \geq 10$ and $np(1-p) \geq 10$) the distribution of our Sample proportions $\hat{p}$ will follow a normal distribution with

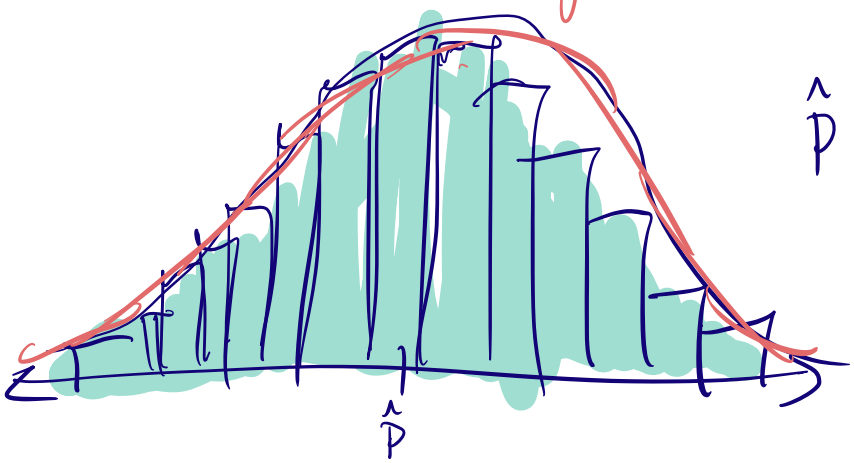
$$\mu = \hat{p} \quad \text{and} \quad SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

where $SE(\hat{p})$ describes the Variability of the distribution of our Sample proportions $\hat{p}$ [Sampling distribution]



plot all proportions as Histograms

(Sampling Distributions)



$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

by CLT

* The Center of The Sampling distribution are always at The population parameter.

* The Variability in our Sampling dist^n decreases as n increases, giving us more precise estimates

* In most cases, we don't Know The population parameter, so $\hat{p}$ is a Reasonable Substitue for p when checking assumptions such as

$$[np \geq 10 \text{ and } np(1-p) \geq 10] \Rightarrow [n\hat{p} \geq 10 \text{ and } n\hat{p}(1-\hat{p}) \geq 10]$$

Similarly, we could plug in $\hat{p}$ to our Standard Error of the Sample proportion

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

§ 5.2 Confidence Interval for a Proportion

- Recall A point estimate is a single number estimate for the population parameter. It is more ideal to give a range of plausible values for the parameter instead with some given confidence, provided a large sample is utilized.

$\hat{p} \longrightarrow p$ (point estimate)

$(LB, UB) \longrightarrow p$ (Interval Estimate)

where LB = point estimate − Margin of Error

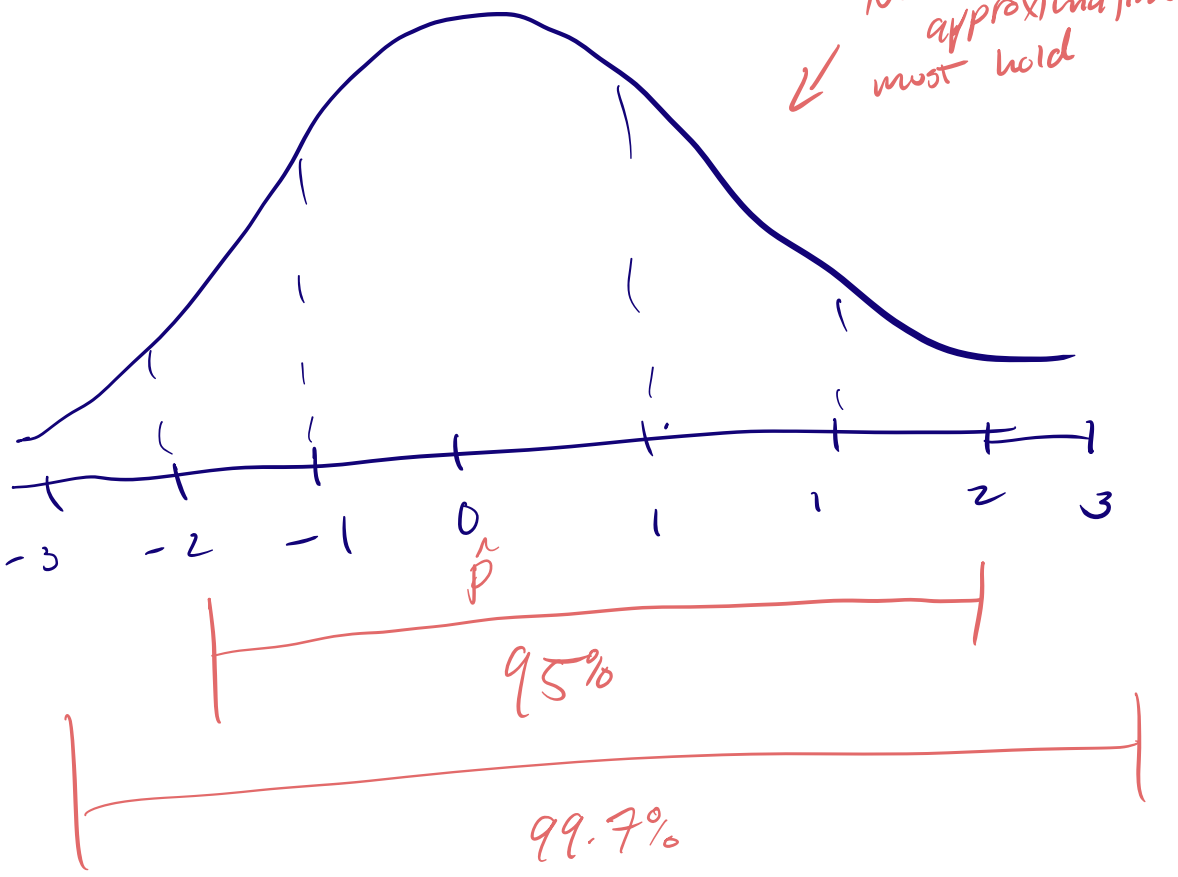
UB = point estimate + Margin of Error

A margin of error consists of the Standard Error and some Confidence Level

Margin of Error = Confidence Level Selected × Standard Error

$ME = z^* SE$

Value is obtained from a standard Normal distn given a certain confidence.



95% extends to 1.96 and −1.96

95% Confidence Interval

$$(LB, UB) = \left(\hat{p} - ME, \hat{p} + ME\right)$$

where $ME = z^* \sqrt{\frac{p(1-p)}{n}}$ and $z^* = 1.96$

All together,

A 95% CI is

$$\left(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}}, \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right)$$

LB UB

* We are 95% Confident The population proportion of "(1)" is Between LB and UB

(1) ↑ statement about population parameter

- Obtain Point Estimate, check normal Assumptions
- Calculate ME
- Calculate LB/UB
- Interpret your Interval Estimate

Show Simulation!

90% CI: use $z^* = 1.65$

95% CI: use $z^* = 1.96$

99% CI: use $z^* = 2.58$

Q: What happens to the Length of the Interval if we increase our confidence?

ex: You work at SunBlock brand A, and your main competitor is Brand B. You test both brands of sunblock on a SRS of 28 subjects. Brand A provides a better UV protection for 18 subjects, but brand B provides better protection for the other 10 subjects.

(a): Show that the assumptions for a large-sample Confidence Interval are satisfied.

- We have SRS, we can assume Sunblock effectiveness for each subject are identically Independently distributed (iid)

- $\hat{p} = \frac{18}{28}$, we can substitute p with $\hat{p}$

and $n\hat{p} = 18$ and $n(1-\hat{p}) = 10$

Since both of these are at least 10, by CLT

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

(b) Construct and Interpret a 95% CI for the proportion of people who get better UV protection from brand A.

- our Sample proportion is $\hat{p} = \frac{18}{28}$ (point estimate)

- $\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.643)(1-0.643)}{28}} = 0.091$$

- With 95% CI level, we have $z^* = 1.96$ since normal approximation holds.

our Margin of Error is

$$ME = z^* SE_{\hat{p}}$$
$$= 1.96 (0.091)$$
$$= 0.177$$

A 95% CI = (LB, UB)

where $LB = \hat{p} - ME = 0.643 - 0.177$

$UB = \hat{p} + ME = 0.643 + 0.177$

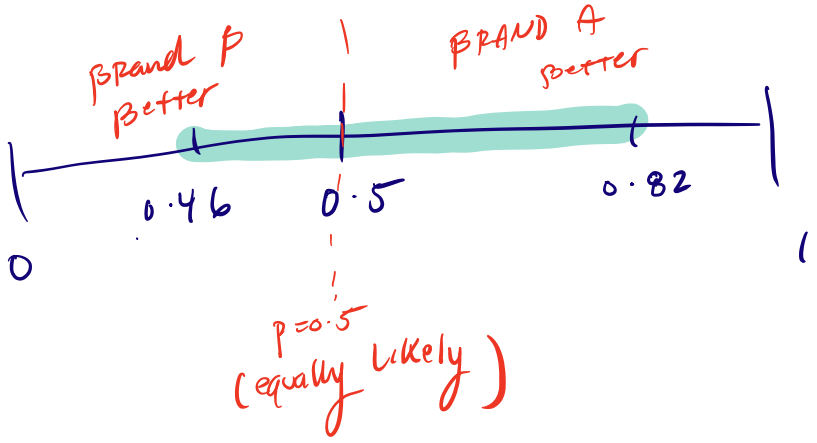
Therefore, 95% CI = $(0.465, 0.820)$

* We are 95% CI That Between 46.5% and 82% of people get better UV protection from Brand A

(c): Given the above Confidence Interval, can we claim that Brand A is better than Brand B?

Nope! If $p > 0.5$ Then brand A provides better protection. However our Interval contains values where $p < 0.5$. Therefore It is plausible for brand B to be better. Similarly our Interval Also contains $p = 0.5$ (Both are the Same)

* Inconclusive Confidence Interval.



§5.3 Hypothesis Test for Proportions

We want to test two competing explanations during an investigation

Two explanations
- Null hypothesis (Assume this to be true)
- Alternative hypothesis (collect data in hopes of supporting this explanation)

Null hypothesis H_0: explanation that claims nothing is happening (no difference)

Alternative hypothesis H_a: explanation that claims something is happening (some difference)

We are interested in trying to "prove" this is TRUE.

Test of significance: Collect and Evaluate Evidence to Decide between the two competing explanations (null vs alternative)

Statement about a population parameter

- Based on our Test of significance, There are two possible outcomes

(1) we fail to Reject The null hypothesis

(2) we Reject The null hypothesis in favor of Alternative hypothesis

* If we fail to Reject our null hypothesis (we did not have enough evidence to support The Alternative hypothesis) That does not imply The null hypothesis is True

---

Recall: you work at Sunblock Brand A. from 28 subjects, Brand A provides better UV protection for 18 subjects, but Brand B provides better protection for the Remaing 10 subjects.

Ex: Translate the real-world claim "people get better UV protection from Brand A than Brand B" into null and Alternative hypotheses (in words and Mathematical Sentences)

H_0: both Brand A and Brand B provide the same UV protection

H_a: Brand A provides better UV protection than Brand B.

Define parameter

Let p = population proportion of individuals for which Brand A was better than Brand B

H_0: $p = 0.5$ ← null value (assume to be true)

H_a: $p > 0.5$ (try to show its true)

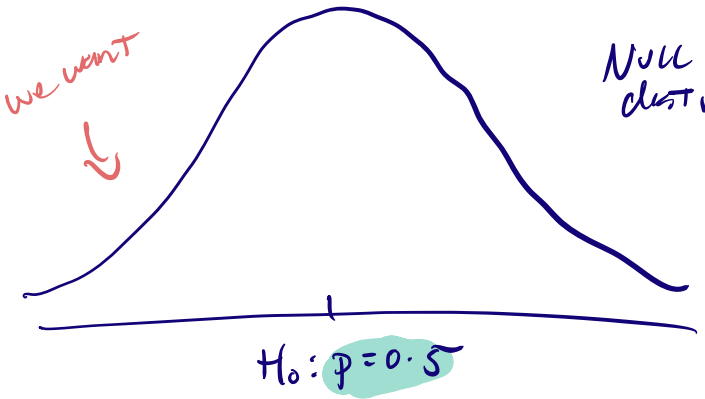
] Specify null / Alternative

Obtain and Evaluate Evidence

- Specify an initial significance level.
(A Threshold to determine if our results are "significant")

We use significance level $= \alpha$, it is common to use $\alpha = 0.05$, but it is not required.

- Check large-sample assumptions for our null distribution



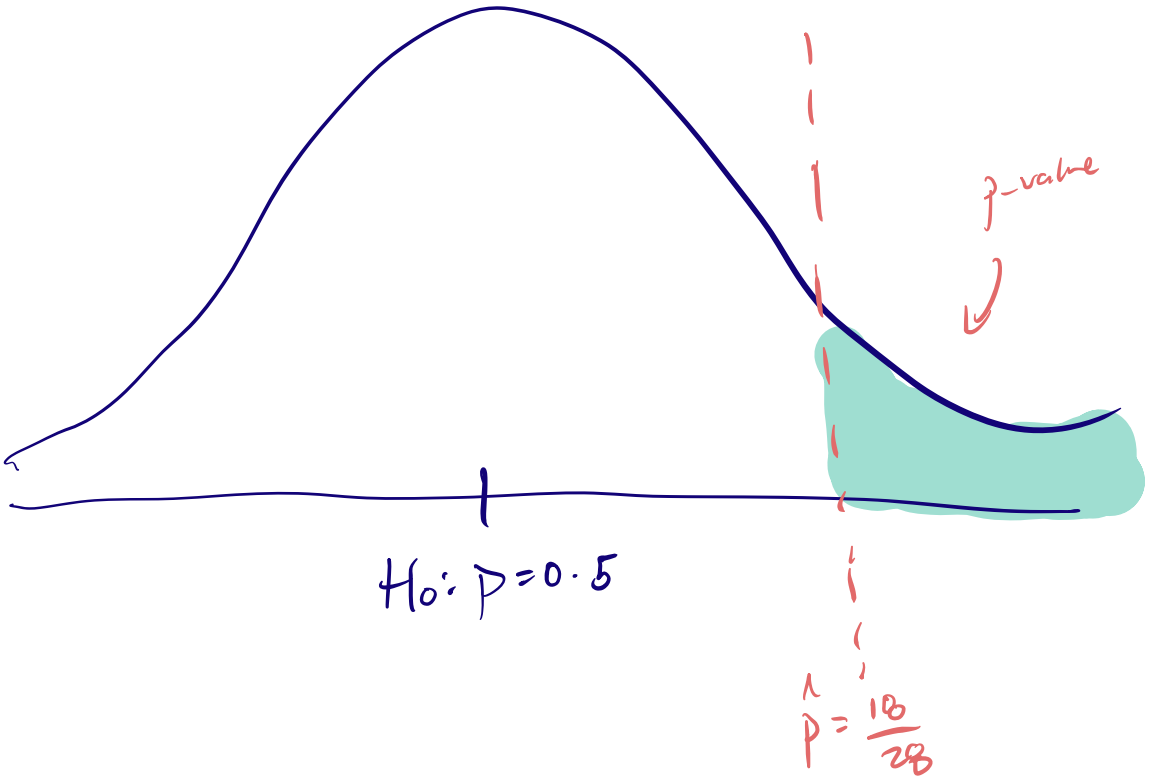
$$\text{Null distn} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

null distribution

$np \geq 10$ and $n(1-p) \geq 10$ using our null value

$28\left(\frac{1}{2}\right) \geq 10$ and $28\left(\frac{1}{2}\right) \geq 10$, since it was SRS independence also holds.

* If the null distribution were true, then our sampling distn should follow a Normal distn with parameter $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ where $p = 0.5$ (our null value)



- Our observed statistic will be used to test the evidence we have in favor of the alternative hypothesis
- we measure such evidence using a p-value.

Defn: A p-value is the probability of observing a test statistic giving us "as much or more evidence in favor of the Alternative hypothesis assuming the null distn is true

In our Example

$$Z = \frac{\text{observed} - \text{null}}{SE}$$

$$Z = \frac{18/20 - 0.5}{\sqrt{\frac{0.5(0.5)}{20}}}$$

$$P(Z \geq z^*) = p\text{-value}$$

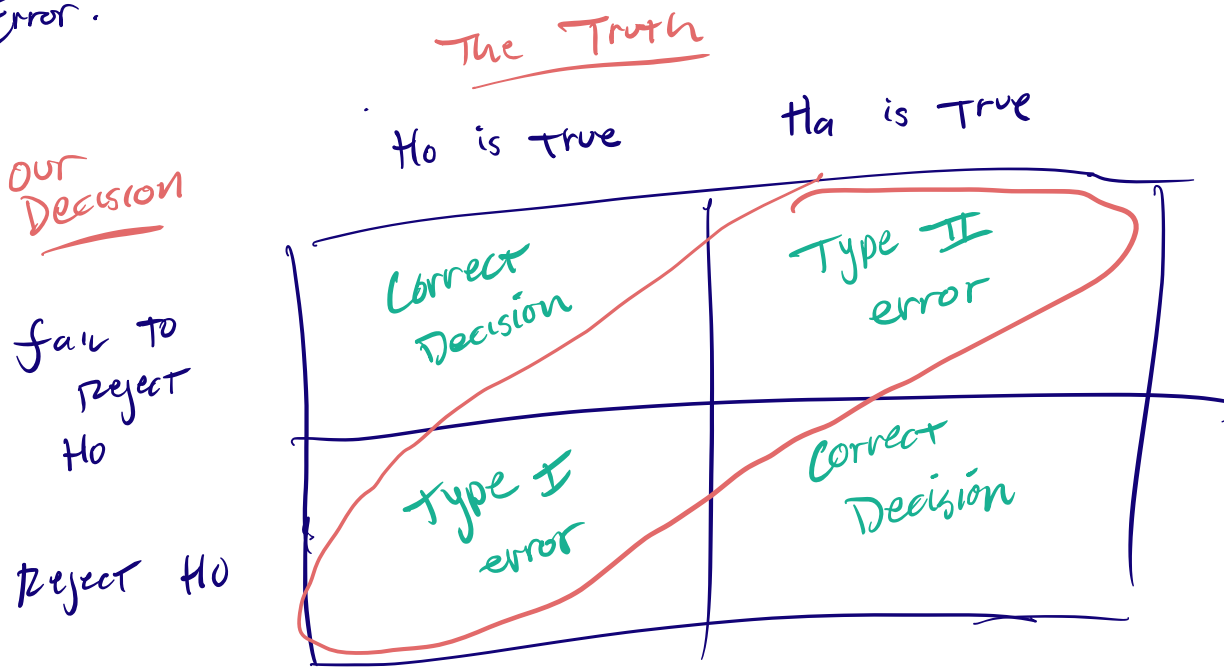
Conclusion:

- If the p-value is less than our initial significance level, we have sufficient evidence in favor of the Alternative Hypothesis and we can Reject our null hypothesis.

- Otherwise, we fail to Reject our null hypothesis if our p-value is greater than our significance level

- Lastly, state Conclusions in the context of the problem.

Based on our conclusion above, we could have made the correct Decision or made an Error.



- Type I error and Type II errors have an inverse relationship. If we minimize Type I error we are increasing Type II error. ↑ chances

- Depending on the context, we have to choose which type of Error we would like to minimize!