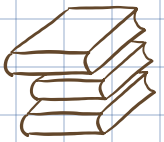
Module Overview

1. **Introduction to Hypothesis Testing**
   Understanding the basic concepts of hypothesis testing, including null hypothesis, alternate hypothesis, P-value, and significance level. etc.
2. **Types of Errors & Framework**
   Exploring the concepts of Type I and Type II errors, and understanding the framework for hypothesis testing.
3. **Numerical vs Categorical Tests**
   Differentiating between hypothesis tests for numerical vs categorical data, and their respective methodologies using different hypothesis tests like Z-tests, T-tests, ANOVA etc.
4. **Categorical vs Categorical Tests**
   Exploring hypothesis tests tailored for comparing categorical variables and analyzing their relationships using Chi- Square test.
5. **Numerical vs Numerical Tests**
   Understanding hypothesis tests designed for comparing numerical variables and assessing their correlations & its types.
6. **Advanced Hypothesis Tests**
   Delving into advanced hypothesis testing techniques such as the Kolmogorov-Smirnov test, A/B testing, and Two-way ANOVA.
7. **Exploratory Data Analysis (EDA) & Feature Engineering**
   Introduction to exploratory data analysis techniques and the process of feature engineering to enhance predictive modeling.
   Techniques like Handling Missing values, Outliers, Categorical & Numerical Data etc.
8. **Yulu Case study**
   You will be working on the real time dataset to extract valuable insights and provide actionable recommendations.
9. **Delhivery Case Study**
   You will be working on the real time dataset to conduct EDA & Feature Engineering to extract valuable insights and provide actionable recommendations.
10. **Module Test & Module Re-test**

Expectations from Learners

- ≥ 85 % PSP
- ≥ 70 % Module Test
- Business case study
- Mock interview



Content and Notes

- GitHub
- Dashboard

} Codes, Handwritten Notes

- End of Module: Cheat Sheet
- Post-Lecture Content in Dashboard

Introduction to Hypothesis Testing

* Motivation: Imagine you are a DS at Amazon. Let's say an executive suggests to remove Add to Cart button

Question: What will be impact of this?

Case 1: No Impact on Sales

Case 2: Sales increased due to better UX

Case 3: Sales decreased

How do we check if Case 2 is true

→ Experiment

→ Hypothesis Testing

→ by collecting Data and evidence

Hypothesis Testing

Consider scenario of Coin Toss in a Match

Default Assumption: Coin is Unbiased/Fair

H (0.5) T (0.5)

Claim: The coin is Biased

What will be needed to say Coin is Biased ?

→ Assumption < Default / Claim

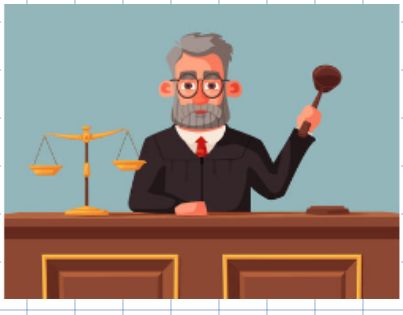
→ Data → Evidence

Let's consider following Exp's to collect Data

	Toss_count	Head_observed	Binomial
1)	10	7	${}^{n}C_{K}\, p^{k}(1-p)^{n-k}$
2)	100	70	${}^{100}C_{70}$ ---
3)	1000	700	${}^{1000}C_{700}$ ---

Key Takeaway: We need to collect enough evidence to Back up the Claim

Judge in Court Example



Consider a judge Hearing case of a Murder Suspect

Default Assumption: →

Innocent until proven Guilty

Claim: Guilty

What data must be collected to reject Default Assumption?

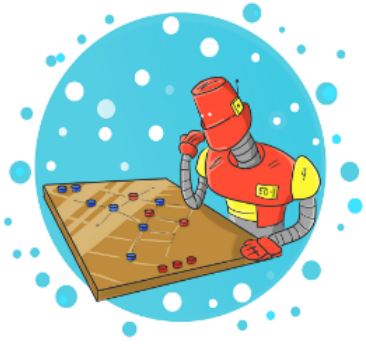
- Fingerprints on murder Scene
- Murder Weapon : Knife
- Forensic Analysis
- Eye-witness
- CCTV Footage

;,

;,

one against the Default Assumption

ML Model Example



Default Assumption
- Old and New Models are Same

Built a new Model

↓

Claim: New model is Better

Note: The Burden of proof is for Rejecting the Default Assumption is on Claim

Hypothesis Testing Terminologies

Let's go back to the judge Scenario

1) Accused is Innocent (Default)
2) Accused is Guilty (Claim)

Default Assumption:

↓

1) Null Hypothesis: (H_0)
   - Status Quo
   - Baseline Assumption
   - Default Assumption

Assumed to be True Until Enough Evidences are present

② Alternate Hypothesis (H_a)

- Claim
- Any assumption other than Null
- It is something that we want to prove by Rejecting H_0

Rejecting H_0 → Success full Rejection / Fail to Reject H_0

P-Value

$P(\text{data} \mid H_0)$

P value is Probability of observing data as extreme as what was Observed under H_0 Null Hypothesis

↓

Innocent

Suppose we have the following data/evidence:-

- **Evidence 1:** Suspect has a knife in their pocket
  - Not enough evidence to say he is guilty
  - Innocent people can also have a knife
- **Evidence 2:** Knife has blood stains on it
  - Not enough evidence.
  - Chef, working with meat / maybe cut himself
- **Evidence 3:**: Blood matches victim
  - Perhaps this is too much to be coincidence
  - But still could be planting of evidence
- **Evidence 4:** Suspect fingerprints found on victim
  - Enough evidence
- **Evidence 5:** CCTV catches suspect at scene of crime / Eyewitness
  - Enough evidence

$P(D_1 \mid H_0) = 0.3$

$P(D_1 \cap D_2 \mid H_0) = 0.07$

$P(D_1 \cap D_2 \cap D_3 \mid H_0) = 0.002$

↓

P — conditional

If we observe all these evidences, then we can reject H_0, because an innocent person is highly unlikely to have this

$P_{condition} \longrightarrow P(Data \mid H_0)$

Key Takeway

> If P-value is very Low
> we Reject H_0

Break Activity

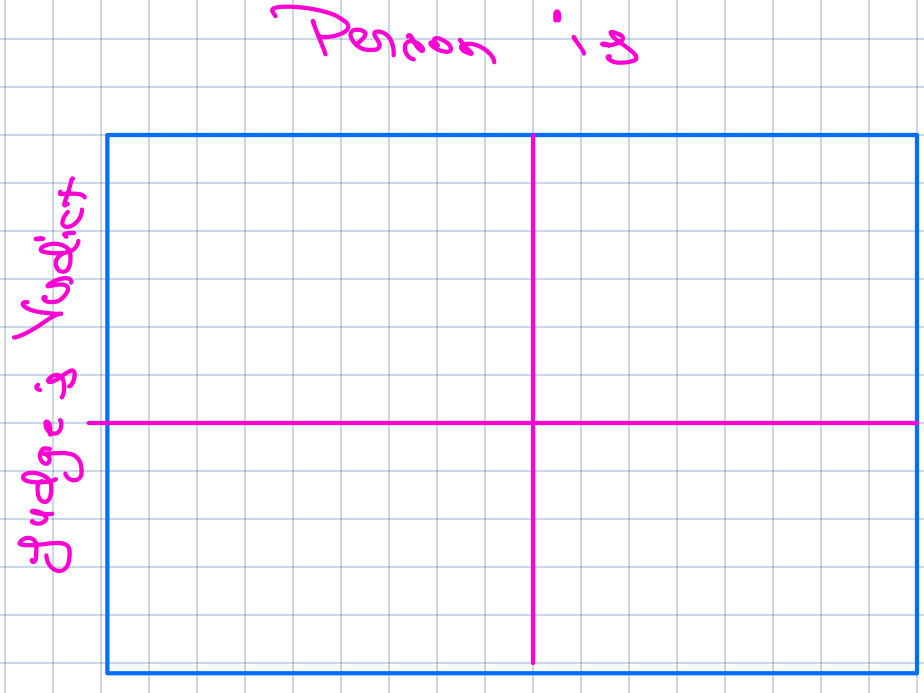
Give $H_0 \longrightarrow$ Innocent

$H_a \longrightarrow$ Guilty

There are 4 Possible Outcome

Can you write All

Hint



$\mu_a < 8$ gms ← Ha

The juice brand claims that the mean sugar content in its juice boxes, produced using a new manufacturing process, is less than 8 grams.

The Food Safety and Standards Authority of India (FSSAI) wants to test the null hypothesis in this scenario. Set up the null (H0) and alternate (H1) hypotheses for this claim, and choose the correct option:

1 user has participated

A	H0: The sugar content in the juice boxes produced using the new manufacturing process is not equal to 8 grams. H1: The sugar content in the juice boxes produced using the new manufacturing process is equal to 8 grams.	0%
B	H0: The sugar content in the juice boxes produced using the new manufacturing process is greater than 8 grams. H1: The sugar content in the juice boxes produced using the new manufacturing process is less than 8 grams.	0%
C	H0: The sugar content in the juice boxes produced using the new manufacturing process is less than 8 grams. H1: The sugar content in the juice boxes produced using the new manufacturing process is greater than 8 grams.	0%
✓ D	H0: The sugar content in the juice boxes produced using the new manufacturing process is greater than or equal to 8 grams. H1: The sugar content in the juice boxes produced using the new manufacturing process is less than 8 grams.	100%

End Quiz Now

H_0 : $\mu \geq 8$ Gm

H_a : $\mu < 8$ Gm

Confidence and Significance Level

- Significance Level (α) → (e.g. 0.05) i.e. a threshold

If P-value $< \alpha$

↓

Reject H_0

Confidence Level $1-\alpha$

0.95

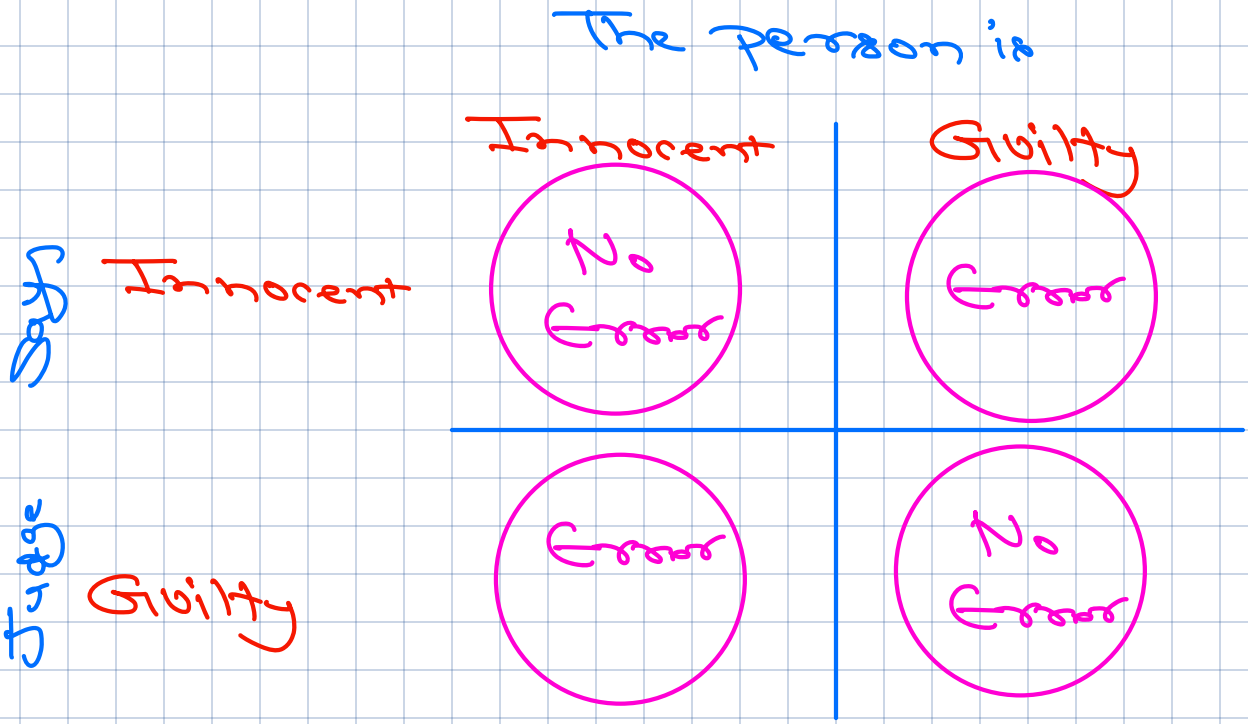
Quiz

P_value → 0.001

α → 0.05

P_value < α → Reject H_0

Types of Errors



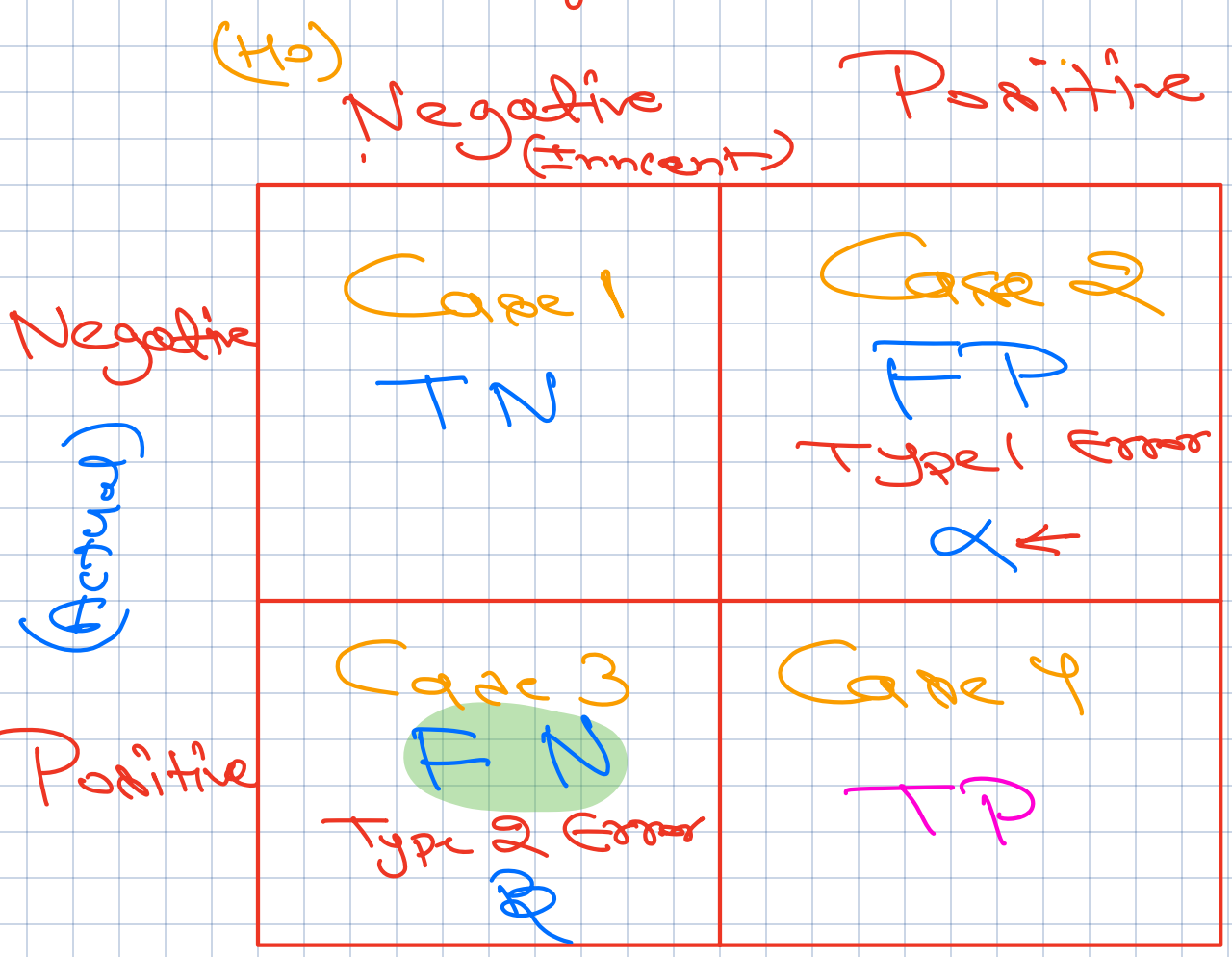
Case 1: We decide that the accused is innocent, and he is actually innocent (i.e. **True Negative**)

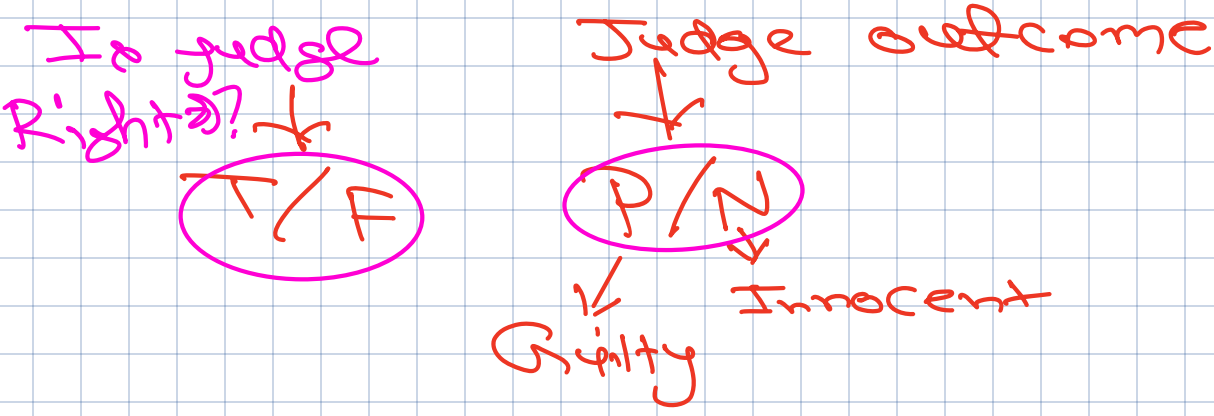
Case 2: We decide that the accused is guilty, but he is actually innocent

Case 3: We decide that the accused is innocent, but he is actually guilty

Case 4: We decide that the accused is guilty, and he is actually guilty (i.e. **True Positive**)

Judge → P but they are F





① Type - I

② α

③ FP

④ X Rejection of Correct H_0

① Type - II

② β

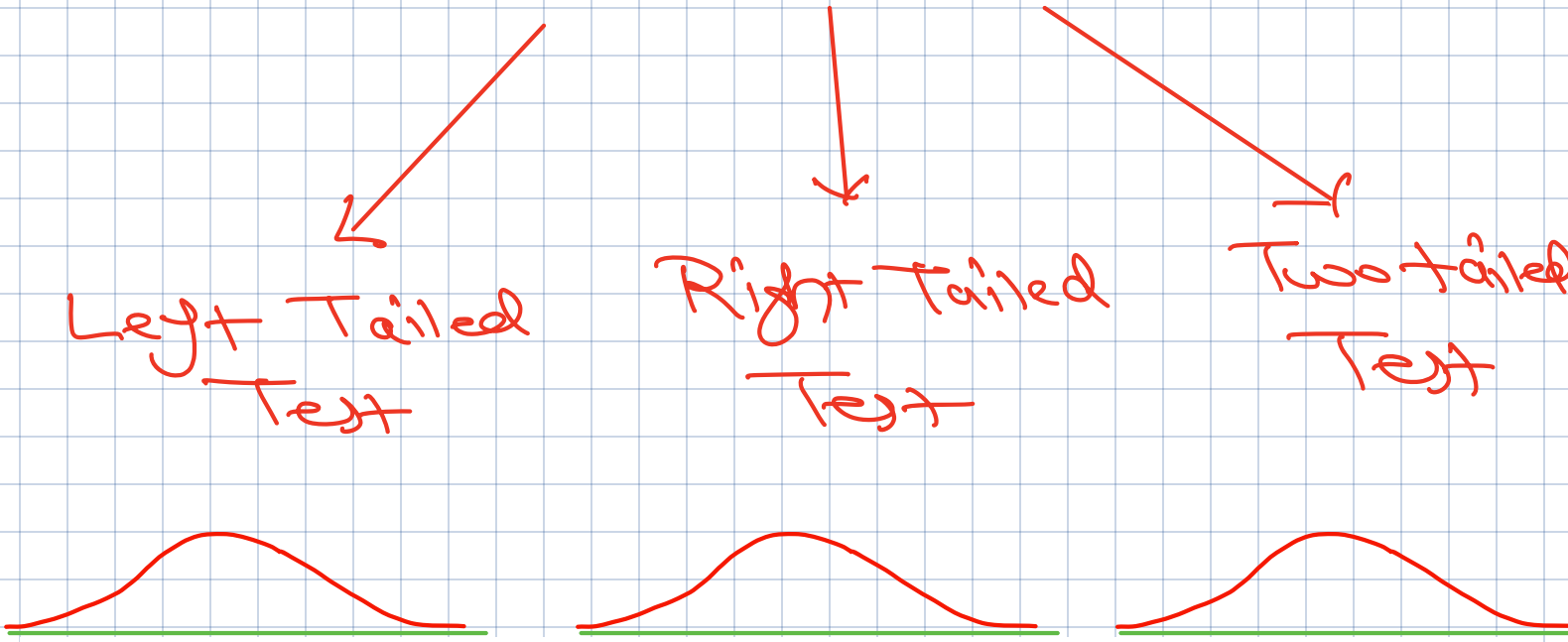
③ FN

④ Failure of Rejection of Incorrect H_0

Confusion Matrix

1. TP
2. TN
3. FP
4. FN

Tailed Tests

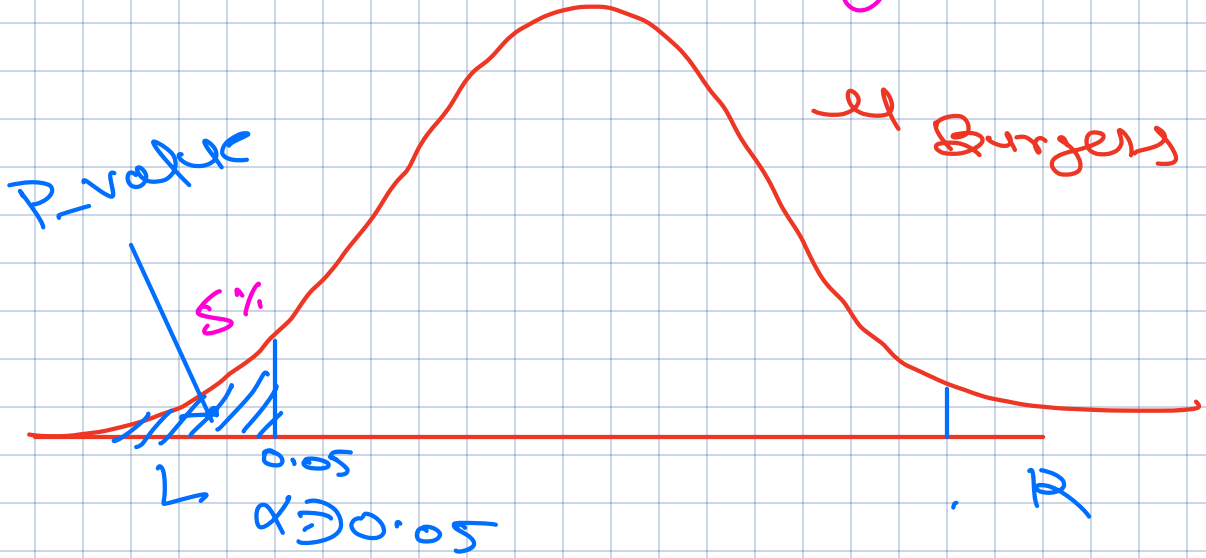


Burger Example

> Suppose there is a burger place that claims that all their burgers are **200 grams**. A customer who consumed their burger is still hungry after eating, and wants to prove that their burgers are lighter, and not as much as promised.

$H_0: \mu \geq 200$

$H_a: \mu < 200$



$\alpha = 0.05$

$CL = 0.95$

Left tailed Test

Consider the example of the Legacy Model, which had an accuracy of 90%. You want to claim that your new model is better.

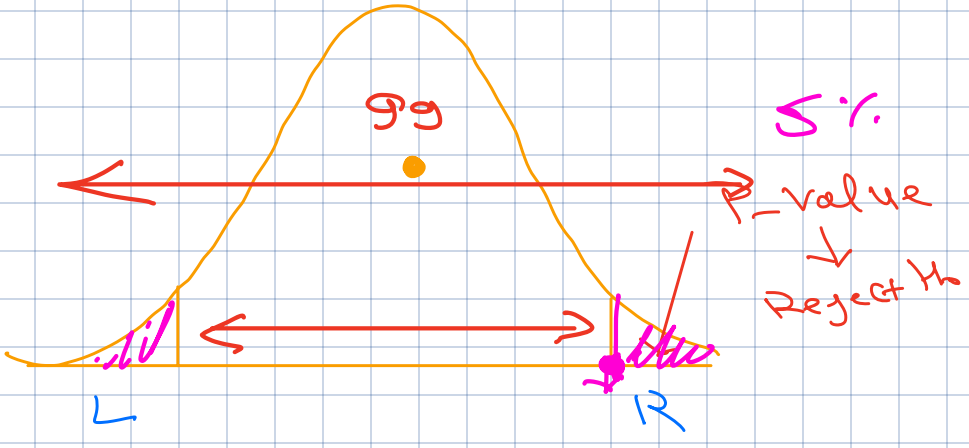
(2) TIL Model

Old Model → 90

New Model > Old Model

$H_0 : \mu_a \leq 90$

$H_a : \mu_a > 90$



Right tailed Test

Suppose you are looking at the height of people in India. It is believed that the average height of Indians is 65 inches. You want to find out if that holds true for the people of your state. Are they taller or shorter?

Case-3: It is believed that Height of Indian People is 65 inch 5'5"

Claim: People of your states are not 65 inches

$H_0 : \mu = 65$

$H_a : \mu \neq 65$

A restaurant says that its new menu item has a mean waiting time of less than equal to 10 minutes.
You claim that the mean wait time is greater than 10 minutes. $\mu > 10$; Ha
What should be the null and alternate hypotheses to test their claim?

3 users have participated

✓	A	H0: μ ≤ 10 and Ha: μ > 10	100%
	B	H0: μ = 10 and Ha: μ > 10	0%
	C	H0: μ = 10 and Ha: μ ≠ 10	0%
	D	H0: μ > 10 and Ha: μ ≤ 10	0%

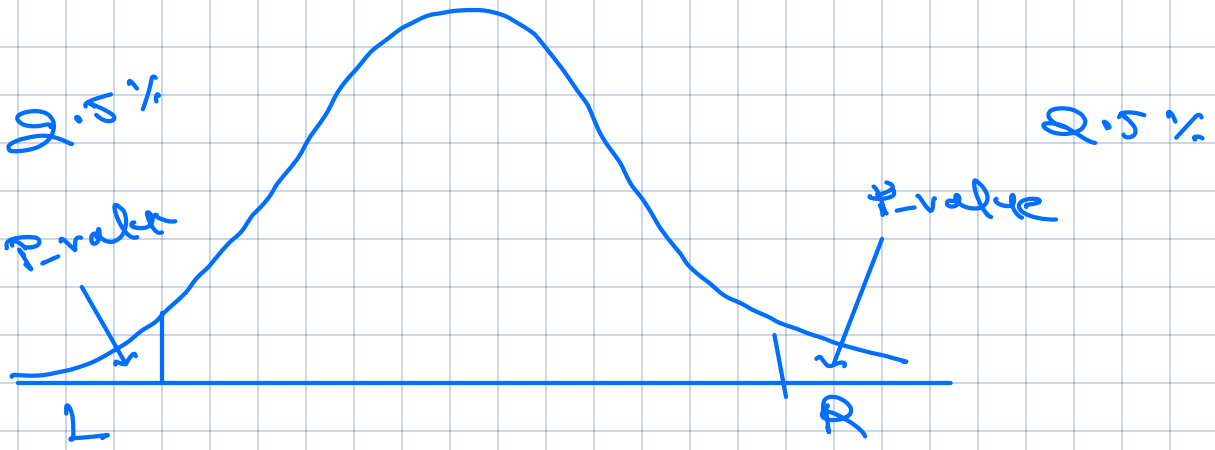
Right

Let's Revisit Coin Toss Example

Steps Hypothesis Testing Framework

1. Setup Null and Alternate Hyp
2. Choose Distribution (Binomial, Normal ..)
3. Select the Tail Type
4. Using Dist: Calculate p-val
5. Compare P-value with α

- Error Types and their relation
- Performing Hypothesis Test



α → 5 % & Two tailed Test

Conclusion

- H_0 → Null Hypothesis
- H_a → Alternate Hypothesis
- α → Significance level
- P-value
- Errors → Type-I → Symbol: FP
  - What does it mean from H_0
  - Type-II Error (FN)