# Hypothesis Testing (VIMP) Part 1

- **IMP for Interviews**

## Hypothesis Testing

- **Hypothesis testing** is a formal procedure used in statistics to decide whether there is enough evidence in a sample of data to infer that a certain condition holds for the entire population.
- It is a **method of making statistical decisions using experimental data.**

### Null Hypothesis ($H_0$)

- The default assumption (e.g., "no effect," "no difference").
- Example: *The new drug has no effect on blood pressure.*
- **Null hypothesis says nothing new is happening.**

> 💡 **Failing to reject the null hypothesis doesn't necessarily mean that the null hypothesis is true;**

### Alternative Hypothesis ($H_1$ or $H_a$)

- The claim we aim to support (e.g., "**there is an effect**").
- Example: *The new drug lowers blood pressure.*

# 1-Tailed vs 2-Tailed test

- In a **two-tailed test**, you check for the possibility of the effect in both directions:
  - The sample mean could be **significantly lower than 13** (one tail), or
  - It could be **significantly higher than 13** (the other tail).
- Because $H_1$ states "not equal" (which covers both possibilities), this test is called **two-tailed**.
- **1-Tailed → < or >**
- **2-Tailed → ≠**

### Significance Level ($\alpha$)

- The probability of rejecting $H_0$ when it is true (**Type I Error**).
- **Common choices:** $\alpha$=0.05 (5%) **or** $\alpha$=0.01 (1%).

### Test Statistic

- A value calculated from sample data (e.g., *t*, z, $\chi^2$) to compare against a critical value or p-value.

# p-value:

💡 **If $p$<$\alpha$, → reject $H_0$**

## p-value calculation in python:

```
p_value = (1 - stats.t.cdf(t_stat, df))
```

- Do `2 *` for 2-tailed test

## Critical value (t*) calculation in python:

```
critical_value = stats.t.ppf(1 - alpha/2, df)
```

## Calculating the t-Statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- **This measures how many standard errors the sample mean is from the hypothesized mean.**

> 💡 If **| $t$ | > critical value**, you **reject $H_o$ (Two-tailed)**

> 💡 **If t > critical value**, you **reject $H_o$ (One-tailed)**

## Type I vs. Type II Errors

- **Type I Error**: Rejecting $H_0$ when it is true (**false positive**).
- **Type II Error**: **Failing to reject $H_0$** when it is false (**false negative**)

## Steps in Hypothesis Testing

1. **State the Hypotheses**:
    - $H_0$ : Null hypothesis.
    - $H_1$: Alternative hypothesis.
2. **Choose Significance Level (α)**:
    - Typically **$\alpha=0.05$**
3. **Select the Appropriate Test**:
    - **Z-test**: For large samples ($n \geq 30$) with known population variance.
    - **t-test**: For small samples ($n<30$) with unknown variance.
    - **Chi-square test**: For categorical data.
    - **ANOVA**: For comparing multiple groups.
4. **Calculate the Test Statistic**:

Example for a **t-test**:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- $\bar{x}$: Sample mean.
- $\mu_0$: Hypothesized population mean.
- $s$: Sample standard deviation.
- $n$: Sample size.

5. **Determine the p-value or Critical Value**:
   - Compare the test statistic to a critical value from tables (e.g., t-table) or calculate the p-value.
6. **Make a Decision**:
   - **Reject $H_0$** if $p<\alpha$ or the test statistic exceeds the critical value.
   - **Fail to reject $H_0$** otherwise.
7. **Interpret the Result**:
   - Example: *There is sufficient evidence to conclude that the new drug lowers blood pressure.*

# Common Tests and When to Use Them

| Test | Use Case | Example |
|---|---|---|
| **One-Sample t-test** | Compare sample mean to a known value. | Is the average height different from 5.8 ft? |
| **Two-Sample t-test** | Compare means of two independent groups. | Do men and women earn different salaries? |
| **Paired t-test** | Compare means of the same group before/after. | Did a training program improve test scores? |
| **Chi-square Test** | Test relationships between categorical variables. | Is there a link between gender and voting preference? |

| | | |
|---|---|---|
| **ANOVA** | Compare means across three or more groups. | Do different fertilizers affect crop yields? |

# Rejection Region Approach/Critical Value Approach

- It involves defining a range of values (the rejection region) for the test statistic.
- If the computed test statistic falls within this region, we reject the null hypothesis ($H_o$); otherwise, we do not reject $H_o$.

1. Formulating hypotheses.
2. Choosing the significance level. (**α=0.05**)
3. Select the Appropriate Test and Compute Degrees of Freedom (e.g., **t-test, z-test**).
    - ***df=n−1***
4. Determine the **Critical Value(s)** (***from t-table or Python***)
    - `t*=stats.t.ppf(1−α/2,df)`
5. Computing the **test statistic**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

6. Determining the critical value.
7. Comparing the test statistic to the critical value.
8. Making a decision (reject or fail to reject $H_o$).

| Test Statistic Approach | P-Value Approach |
|---|---|
| State $H_0$ and $H_1$ | State $H_0$ and $H_1$ |
| Determine test size $\alpha$ and find the critical value (CV) | Determine test size $\alpha$ |
| Compute a test statistic (TS) | Compute a test statistic and its p-value |
| Reject $H_0$ if TS > CV | Reject $H_0$ if p-value < $\alpha$ |
| Substantive interpretation | Substantive interpretation |

| Step | Test Statistic Approach | P-Value Approach | Confidence Interval Approach |
|---|---|---|---|
| 1 | State $H_0$ and $H_1$ | State $H_0$ and $H_1$ | State $H_0$ and $H_1$ |
| 2 | Determine test size $\alpha$ and find the critical value (CV) | Determine test size $\alpha$ | Determine test size $\alpha$ or 1-$\alpha$, and a hypothesized value |
| 3 | Compute a test statistic (TS) | Compute a test statistic and its p-value | Construct the (1-$\alpha$)100% confidence interval (CI) |
| 4 | Reject $H_0$ if TS > CV | Reject $H_0$ if p-value < $\alpha$ | Reject $H_0$ if a hypothesized value does not exist in CI |
| 5 | Substantive interpretation | Substantive interpretation | Substantive interpretation |

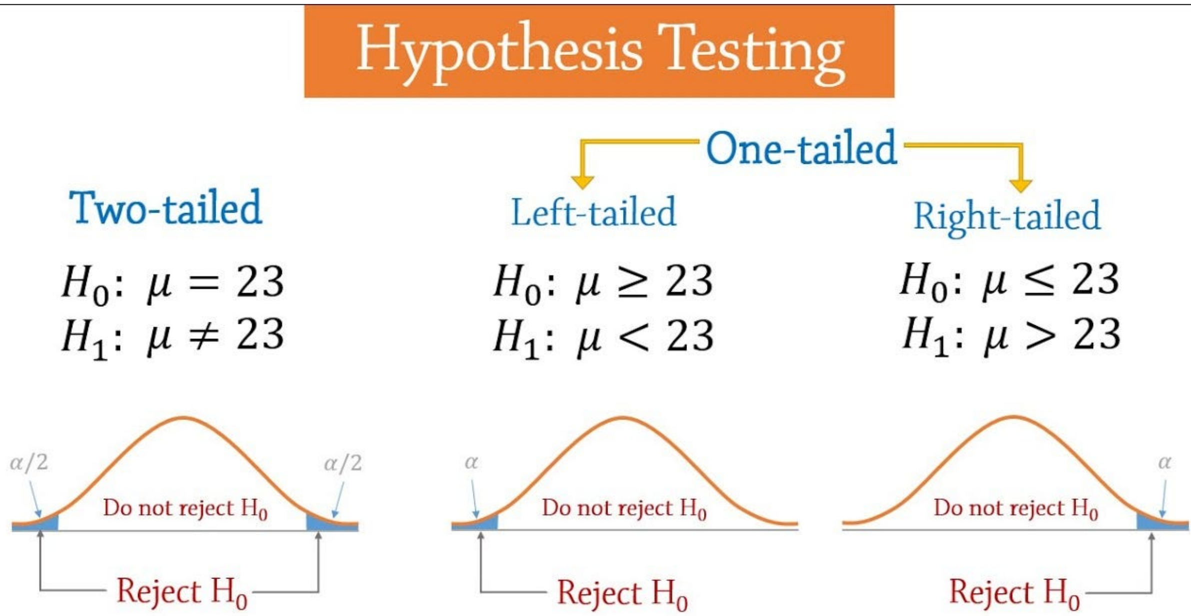
```
                  [t-Distribution Curve]

    |<------- Rejection Region ------>|<--- Acceptance Reg
  -t*                                  0
    |                                   |
    |                                   |
    +---------+                         |
              |                         |
         (t_statistic)           (Test Statistic)
              |                         |
              V                         V

------------------------------------------------------------
|             Rejection Region Approach
|   Decision: If t_statistic < -t* or t_statistic > t*,
```

```
    |           then reject Hₒ. Otherwise, do not reject Hₒ.
    -----------------------------------------------------------


    -----------------------------------------------------------
    |               p-value Approach
    |   Compute the p-value = 2 * P(T ≥ |t_statistic|)
    |   Decision: If p-value < α, reject Hₒ; else, do not rejec
    -----------------------------------------------------------
```




# EXAMPLE

*Q. Suppose a company is evaluating the impact of a new training program on the productivity of its employees. The company has data on the average productivity of its employees before implementing the training program.*

*The average productivity was **50 units per day with a known population standard deviation of 5 units**. After implementing the training program, the company measures the productivity of a random sample of **30 employees**. The* *sample has an **average productivity of 53 units per day.** The company wants to know if the new training program has significantly increased productivity.*

Here,

$H_0$=μ=50

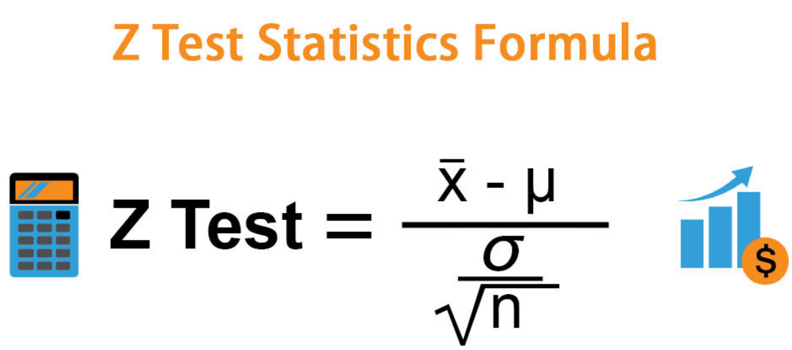
σ = 5

α = 0.05 → 95%

n = 30

x̄ = 53

- **We have population SD, therefore we'll use <u>Z-test</u>**




Z= $(53-50)/(5/\sqrt{30})$ = 3.28

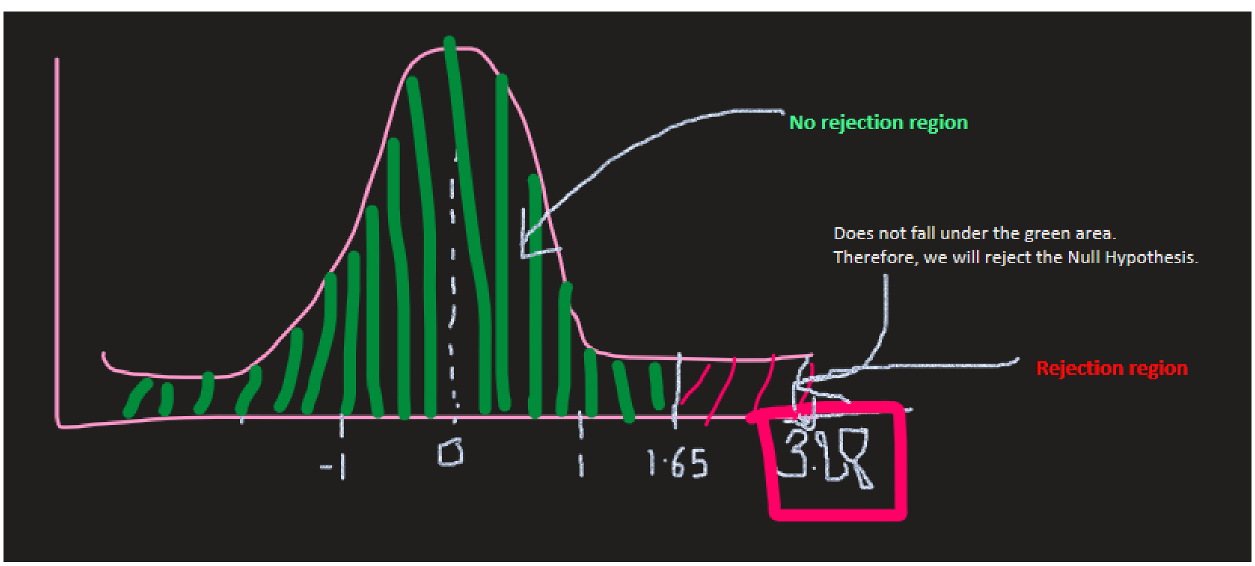
### Z Critical values for 95% Ci = 1.65

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

| Z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | .50000 | .50399 | .50798 | .51197 | .51595 | .51994 | .52392 | .52790 | .53188 | .53586 |
| 0.1 | .53983 | .54380 | .54776 | .55172 | .55567 | .55962 | .56356 | .56749 | .57142 | .57535 |
| 0.2 | .57926 | .58317 | .58706 | .59095 | .59483 | .59871 | .60257 | .60642 | .61026 | .61409 |
| 0.3 | .61791 | .62172 | .62552 | .62930 | .63307 | .63683 | .64058 | .64431 | .64803 | .65173 |
| 0.4 | .65542 | .65910 | .66276 | .66640 | .67003 | .67364 | .67724 | .68082 | .68439 | .68793 |
| 0.5 | .69146 | .69497 | .69847 | .70194 | .70540 | .70884 | .71226 | .71566 | .71904 | .72240 |
| 0.6 | .72575 | .72907 | .73237 | .73565 | .73891 | .74215 | .74537 | .74857 | .75175 | .75490 |
| 0.7 | .75804 | .76115 | .76424 | .76730 | .77035 | .77337 | .77637 | .77935 | .78230 | .78524 |
| 0.8 | .78814 | .79103 | .79389 | .79673 | .79955 | .80234 | .80511 | .80785 | .81057 | .81327 |
| 0.9 | .81594 | .81859 | .82121 | .82381 | .82639 | .82894 | .83147 | .83398 | .83646 | .83891 |
| 1.0 | .84134 | .84375 | .84614 | .84849 | .85083 | .85314 | .85543 | .85769 | .85993 | .86214 |
| 1.1 | .86433 | .86650 | .86864 | .87076 | .87286 | .87493 | .87698 | .87900 | .88100 | .88298 |
| 1.2 | .88493 | .88686 | .88877 | .89065 | .89251 | .89435 | .89617 | .89796 | .89973 | .90147 |
| 1.3 | .90320 | .90490 | .90658 | .90824 | .90988 | .91149 | .91309 | .91466 | .91621 | .91774 |
| 1.4 | .91924 | .92073 | .92220 | .92364 | .92507 | .92647 | .92785 | .92922 | .93056 | .93189 |
| 1.5 | .93319 | .93448 | .93574 | .93699 | .93822 | .93943 | .94062 | .94179 | .94295 | .94408 |
| 1.6 | .94520 | .94630 | .94738 | .94845 | .94950 | .95053 | .95154 | .95254 | .95352 | .95449 |
| 1.7 | .95543 | .95637 | .95728 | .95818 | .95907 | .95994 | .96080 | .96164 | .96246 | .96327 |
| 1.8 | .96407 | .96485 | .96562 | .96638 | .96712 | .96784 | .96856 | .96926 | .96995 | .97062 |
| 1.9 | .97128 | .97193 | .97257 | .97320 | .97381 | .97441 | .97500 | .97558 | .97615 | .97670 |
| 2.0 | .97725 | .97778 | .97831 | .97882 | .97932 | .97982 | .98030 | .98077 | .98124 | .98169 |
| 2.1 | .98214 | .98257 | .98300 | .98341 | .98382 | .98422 | .98461 | .98500 | .98537 | .98574 |
| 2.2 | .98610 | .98645 | .98679 | .98713 | .98745 | .98778 | .98809 | .98840 | .98870 | .98899 |

- As it's a 1-tailed test, we didn't divide it by 2
  - Otherwise, for 2-tailed, z would have been 1.96

3.28 > 1.65 ... Therefore, we **reject the $H_0$**



***Q. Suppose a snack food company claims that their Lays wafer packets contain an average weight of 50 grams per packet. To verify this claim, a consumer watchdog organization decides to test a random sample of Lays wafer packets. The organization wants to determine whether the actual average weight differs significantly from the claimed 50 grams. The organization collects a random sample of 40 Lays wafer packets and measures their weights. They find that the sample has an average weight of 49 grams, with a known population standard deviation of 4 grams.***

$\mu = 50 \quad n = 40 \quad \bar{x} = 49 \quad \sigma = 4$

1) $H_0: \mu = 50 \quad H_a: \mu \neq 50$

2) $\alpha = 0.05$

3) Normality ✓ $\sigma$ ✓ → Z test

4) Z test

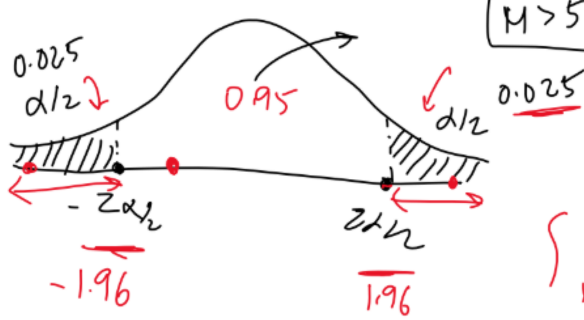
$\alpha = 5\%$

5) Z

6) $Z = \dfrac{49 - 50}{4/\sqrt{40}} = \dfrac{-\sqrt{40}}{4} = \boxed{-1.58}$

$\boxed{\mu > 50}$ $\quad \mu \neq 50$

can't reject the NULL hypothesis




$\mu \neq 50$

$\left\{ \begin{array}{l} \mu > 50 \\ \mu < 50 \end{array} \right\}$

👆 **THIS APPROACH IS NOT USUALLY PREFERRED BECAUSE IT DOES NOT TELL YOU THE STRENGTH OF THE REJECTION.**

- i.e. It cannot differentiate between z=2 & z=15
- Therefore, we use **p-value approach**

**With p-value approach:**

- We calculate the p-value
- We can measure strength

# Type 1 vs Type 2 Error- VVVIMP for Interviews

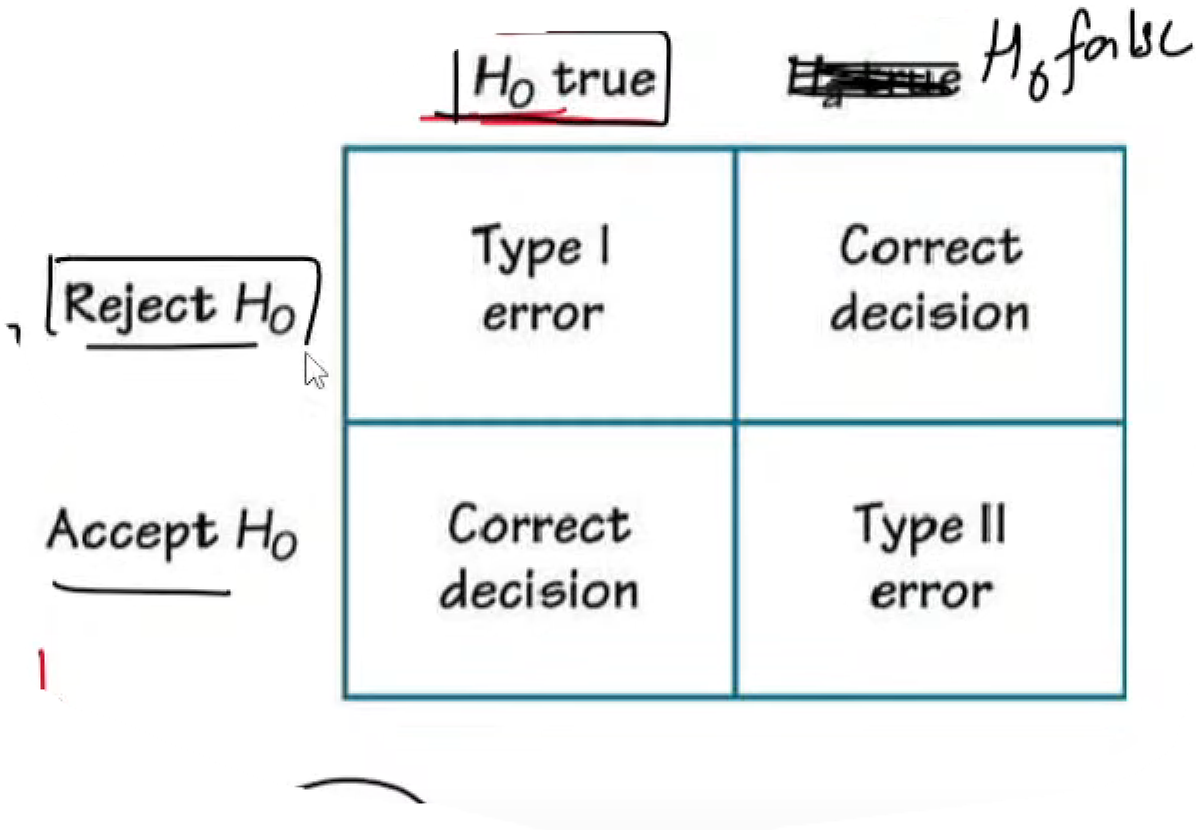
## Type I Error (False Positive)

- **Definition:** Occurs when you **reject the null hypothesis ($H_0$)** when it is actually **true**.
- **FALSE ALARM**
- **Interpretation:** You claim there is an effect or difference when, in fact, there isn't one.
- The probability of committing a Type I error is denoted by **α (alpha)**, which is also the significance level of the test.
- **Example:** In a medical trial, concluding that a new drug is effective when it actually has no effect is a Type I error.
- Set a lower significance level (α), e.g., $\alpha=0.01$ to reduce the Type-1 Error.

## Type II Error (False Negative)

- **Definition:**Occurs when you **fail to reject the null hypothesis ($H_0$)** when the alternative hypothesis ($H_1$) is actually **true**.

- **Interpretation:**You claim there is no effect or difference when, in fact, there is one.
- **Notation:** The probability of committing a Type II error is denoted by **β (beta)**.
    - **The power of the test is 1−β, which indicates the probability of correctly detecting an effect.**
    - **We can decrease Type-2 error by increasing power of the test.**
- **Example:**In the same medical trial, concluding that the new drug is not effective when it actually works is a Type II error.




| Actual Truth | Decision Made | Error Type |
| --- | --- | --- |
| $H_0$ is True | Reject $H_0$ | **Type I Error (α)** |
| $H_0$ is True | Fail to Reject $H_0$ | Correct Decision |
| $H_0$ is False ($H_1$ True) | Reject $H_0$ | Correct Decision |

| Actual Truth | Decision Made | Error Type |
|---|---|---|
| $H_0$ is False ($H_1$ True) | Fail to Reject $H_0$ | **Type II Error (β)** |

Ex.

H0 = Person is innocent.

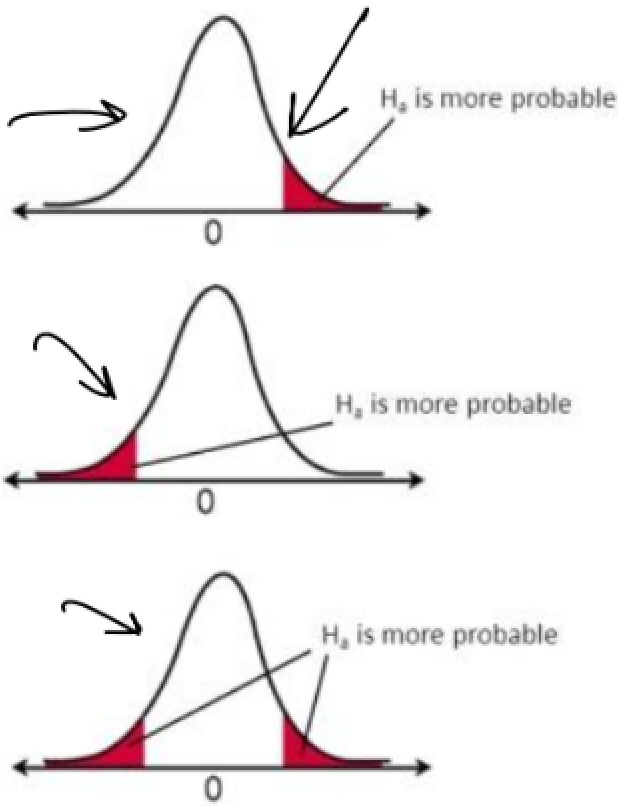
H1= Person is not innocent.

**Type-1 Error**→ The person is innocent. But still they are being punished.

**Type-2 Error**→ We are acquitting the person. But that person is not innocent.

# 1-Tailed vs 2-Tailed test

- In a **two-tailed test**, you check for the possibility of the effect in both directions:
    - The sample mean could be **significantly lower than 13** (one tail), or
    - It could be **significantly higher than 13** (the other tail).
- Because $H_1$ states "not equal" (which covers both possibilities), this test is called **two-tailed**.
- **1-Tailed** → **< or >**
- **2-Tailed** → **≠**




# Where can be Hypothesis Testing Applied?

## Medicine and Healthcare

- Test the effectiveness of a new drug or treatment.
- Compare patient outcomes between different therapies.

## Business and Marketing

- Evaluate the impact of a new marketing campaign.
- Test whether a new product feature increases customer satisfaction.

## Social Sciences

- Determine if there's a relationship between education level and income.

- Test the effectiveness of a new teaching method.

### Quality Control

- Check if a manufacturing process meets quality standards.
- Compare the performance of two production lines.

### Finance

- Test if a new investment strategy yields higher returns.
- Compare the risk profiles of different portfolios.

# Machine Learning Applications of Hypothesis Testing

## A. Model Evaluation

1. **Compare Model Performance**:
    - Test if one model (e.g., Random Forest) performs significantly better than another (e.g., Logistic Regression) on a dataset.
    - Example: Use a **paired t-test** to compare cross-validation scores of two models.
2. **Statistical Significance of Metrics**:
    - Test if the accuracy, precision, or recall of a model is significantly better than a baseline.
    - Example: Use a **z-test** to compare classification accuracy to a random guess.

## B. Feature Selection

1. **Test Feature Importance**:
    - Determine if a feature significantly contributes to the model's performance.

- Example: Use a **chi-square test** for categorical features or **ANOVA** for continuous features.

2. **Correlation Testing**:
    - Test if a feature is significantly correlated with the target variable.
    - Example: Use **Pearson's correlation test** for linear relationships.

## C. A/B Testing in ML Systems

1. **Model Deployment**:
    - Test if a new model performs better than the current one in production.
    - Example: Use a **two-sample t-test** to compare key metrics (e.g., click-through rates) between the two models.
2. **Hyperparameter Tuning**:
    - Test if a specific hyperparameter setting leads to significantly better performance.
    - Example: Use a **paired t-test** to compare validation scores across different hyperparameter configurations.

## D. Data Drift Detection

1. **Test for Distribution Shifts**:
    - Detect if the input data distribution has changed over time (e.g., due to concept drift).
    - Example: Use the **Kolmogorov-Smirnov test** to compare training and test data distributions.
2. **Feature Drift**:
    - Test if the distribution of a specific feature has changed.
    - Example: Use a **chi-square test** for categorical features or **t-test** for continuous features.

---

## E. Bias and Fairness Testing

1. **Test for Bias**:
    - Check if a model's predictions are biased against a specific group (e.g., gender, race).
    - Example: Use a **chi-square test** to compare prediction outcomes across groups.
2. **Fairness Metrics**:
    - Test if fairness metrics (e.g., equal opportunity, demographic parity) are significantly different across groups.
    - Example: Use a **t-test** to compare fairness metrics.

---

### F. Anomaly Detection

1. **Test for Outliers**:
    - Determine if a data point is significantly different from the rest.
    - Example: Use a **Grubbs' test** or **Z-score test** for outlier detection.

---

### G. Hypothesis Testing in Deep Learning

1. **Test for Overfitting**:
    - Compare training and validation performance to detect overfitting.
    - Example: Use a **paired t-test** to compare training and validation losses.
2. **Test for Model Stability**:
    - Check if a model's performance is consistent across different random seeds.
    - Example: Use an **ANOVA test** to compare performance across multiple runs.

## Key Hypothesis Tests in ML

| Test | Use Case | Example |
|---|---|---|

| | | |
|---|---|---|
| **t-test** | Compare means of two groups. | Compare model accuracy on two datasets. |
| **ANOVA** | Compare means of three or more groups. | Compare performance of multiple models. |
| **Chi-square Test** | Test relationships between categorical variables. | Test if a feature is independent of the target. |
| **Kolmogorov-Smirnov** | Compare two distributions. | Detect data drift between training and test sets. |
| **Mann-Whitney U Test** | Compare medians of two groups (non-parametric). | Compare model performance on skewed data. |

```
      +---------------------------------------+
      |       Machine Learning Applications  |
      +---------------------------------------+
                 /           \\
                /             \\
      +---------+               +----------+
      | Feature |               | Model    |
      |Selection|               |Comparison|
      +---------+               +----------+
           |                          |
           |                          |
+-----------------+          +---------------------+
| Test if Feature|          | Test if performance|
| is significant |          | improvements are   |
| (t-test, chi-  |          | statistically      |
| square, etc.)  |          | significant (A/B   |
+-----------------+          | testing, etc.)     |
           |                          |
      +---------+               +----------+
      | Validate|               | Detect   |
      |Assumptions|             | Drift (Feature/    |
      | (Normality,|            | Concept Drift)    |
```

```
        | Equal Variances)|       +----------+
        +---------+
```