Introduction to Hypothesis

A **hypothesis** is a formal statement or assumption made about a population parameter. It is typically tested using sample data. Hypothesis testing is one of the principal instruments in research, especially in inferential statistics.

A **statistical hypothesis** is a conjecture or assumption about one or more parameters of a population, which is tested through analysis of sample data.

Objectives of Hypothesis Testing

- To assess whether the sample data supports a given assumption about the population.
- To make **probabilistic statements** about the population parameters.
- To test the **significance** of relationships and differences between variables.

Characteristics of a Good Hypothesis

A hypothesis should:

1. Be **clear and precise**.
2. Be **testable** through data.
3. State a **relationship between variables**.
4. Be **specific and limited in scope**.
5. Be **simple and understandable**.
6. Be **testable within a reasonable time**.
7. Have **empirical reference** (i.e., it should be based on observable phenomena).

Types of Hypotheses

A. Research Hypothesis

- A **tentative assumption** or proposition about a research problem.
- It motivates investigation and provides direction for data collection.
- Example: *"Increased social media use reduces academic performance."*

B. Statistical Hypothesis

- A **formal and testable** statement about a population parameter.
- Assessed using statistical techniques and sample data.
- Typically tested through **hypothesis testing procedures**.

Types of Statistical Hypotheses

A. Null Hypothesis (H_0)

- Denotes **no effect, no difference**, or no relationship.
- It is the **default assumption** that the researcher tries to test or disprove.
- Example:
  H_0: $\mu = 100$ → The population mean is equal to 100.

B. Alternative Hypothesis (H_1 or Ha)

- Represents the **opposite** of the null hypothesis.
- Suggests the **presence** of an effect, difference, or relationship.
- Can be one-sided or two-sided:
  - H_1: $\mu \neq 100$ (two-tailed)
  - H_1: $\mu > 100$ or H_1: $\mu < 100$ (one-tailed)

Process of Hypothesis Testing

1. **Formulate the hypotheses**: Define H_0 and H_1.
2. **Select significance level (α)**: Common values are 0.05, 0.01.
3. **Choose the appropriate test**: z-test, t-test, F-test, etc.
4. **Determine the test statistic**: Based on sample data.
5. **Define the critical region**: Using statistical tables.
6. **Compute the value of the test statistic**.
7. **Make a decision**:
   - If the test statistic falls in the critical region, **reject H_0**.
   - If it does not, **fail to reject H_0**.

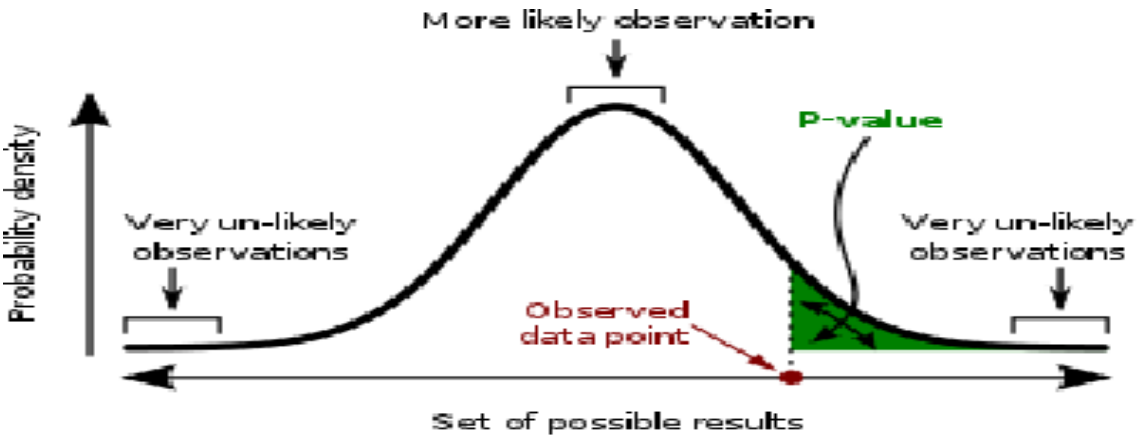
Level of Significance: The probability level below which we reject the hypothesis is called level of significance. The levels of significance usually employed in testing of hypothesis are 5% and 1%.

P-Value: The p-value is the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected.

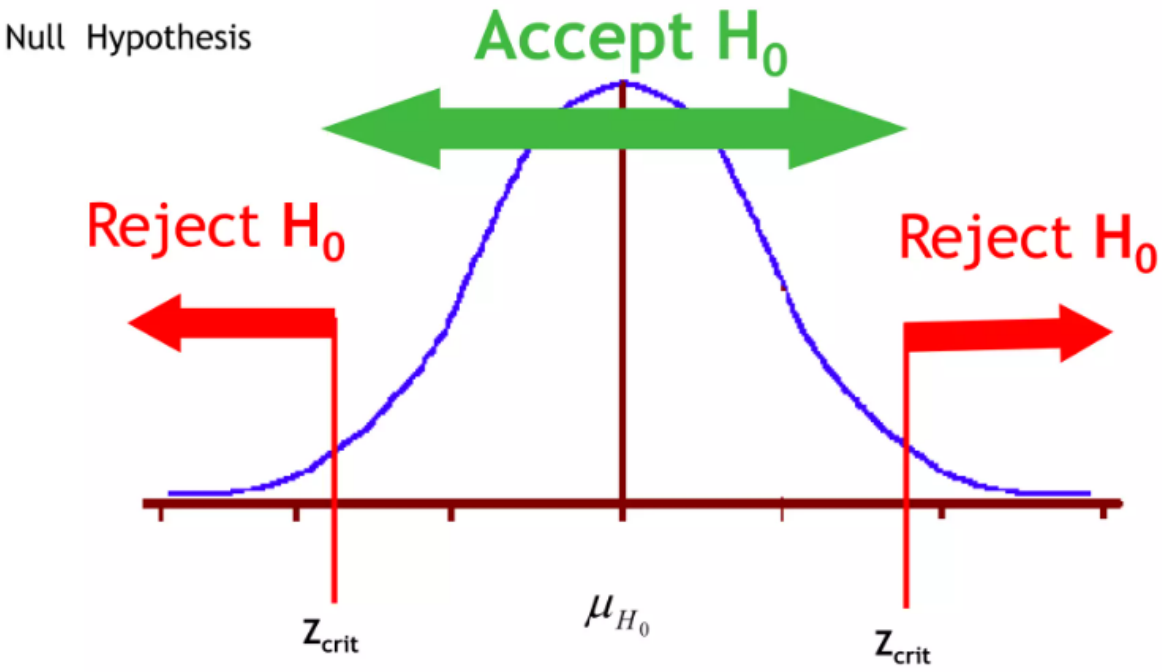
A p value is used in hypothesis testing to help you support or reject the null hypothesis. The p value is the evidence against a null hypothesis. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

P values are expressed as decimals although it may be easier to understand what they are if you convert them to a percentage. For example, a p value of 0.0254 is 2.54%. This means there is a 2.54% chance your results could be random (i.e. happened by chance). That's pretty

tiny. On the other hand, a large p-value of .9(90%) means your results have a 90% probability of being completely random and not due to anything in your experiment. Therefore, the smaller the p-value, the more important ("significant") your results.



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.



Decision rule or Test of Hypothesis: A decision rule is a procedure that the researcher uses to decide whether to accept or reject the null hypothesis. The decision rule is a statement that tells under what circumstances to reject the null hypothesis. The decision rule is based on specific values of the test statistic (e.g., reject H0 if Calculated value > table value at the same level of significance)

1. **Types of Error:** In the context of testing of hypotheses, there are basically two types of errors we can make.

a. **Type 1 error:** To reject the null hypothesis when it is true is to make what is known as a type I error. The level at which a result is declared significant is known as the type I error rate, often denoted by α.
b. **Type II error:** If we do not reject the null hypothesis when in fact there is a difference between the groups, we make what is known as a type II error. The type II error rate is often denoted as β.

In a tabular form the said two errors can be presented as follows:

Particulars	Decision	
	Accept H0	**Reject H0**
H0 (True)	Correct Decision	Type I error (α error)
H0 (False)	Type II error (β error)	Correct decision

Z-Test

Purpose and Applications

The Z-test is used to test hypotheses about population means when the sample size is large **(n>30)** and the population variance is known.

It is commonly applied in scenarios where you want to compare a sample mean to a known population mean or test proportions in large samples.

A z-test is based on the **standard normal distribution (Z-distribution)**. It is used to test hypotheses about population means and proportions when:

- Sample size is **large (n > 30)**, or
- Population variance is **known**.

Applications:

- Testing **sample mean** vs. population mean.
- Comparing **two sample means** (large samples).
- Testing **sample proportion** vs. population proportion.
- Comparing **two proportions** from independent large samples.

Assumptions

- **Normal Distribution**: The data should follow a normal distribution.
- **Known Population Variance**: The variance of the population must be known.

If the calculated Z-score exceeds the critical value from the standard normal distribution (e.g., $Z = 1.96$ for a two-tailed test at $\alpha=0.05\alpha=0.05$), you reject the null hypothesis that the sample mean is equal to the population mean.

Advantages:

- Simple and quick to use for large datasets.
- Effective when population parameters are known.

t-Test

Purpose and Applications

The t-test is used for comparing means when the sample size is small ($n<30$) or the population variance is unknown.

The t-test is based on the **t-distribution**, used when:

- Sample size is **small ($n \leq 30$)**, and
- Population variance is **unknown**.

Types of t-tests:

1. **One-sample t-test**:
   Tests whether the sample mean differs from a known or hypothesized population mean.
2. **Independent two-sample t-test**:
   Tests whether the means of **two independent groups** are significantly different.
3. **Paired t-test**:
   Used when the samples are **related or matched** (e.g., before and after tests on the same subject).

Assumptions

- **Normal Distribution**: The data should follow a normal distribution.
- **Homogeneity of Variances**: For independent samples, the variances should be equal (though this can be relaxed with certain modifications).

If the calculated t-score exceeds the critical value from the t-distribution table (which depends on the degrees of freedom and chosen significance level), you reject the null hypothesis that the means are equal.

F-Test (ANOVA)

Purpose and Applications

The F-test, often used in the context of Analysis of Variance (ANOVA),

is used to compare variances across multiple groups or test differences between group means. It is particularly useful for comparing means across three or more groups.

An F-test is used to compare **two variances** or to perform **Analysis of Variance (ANOVA)**, which evaluates **differences between more than two means**.

Types of F-tests:

1. **Test of equality of variances**:
   Checks whether two populations have equal variances.
2. **One-way ANOVA**:
   Compares the means of three or more independent groups based on one factor.
3. **Two-way ANOVA**:
   Compares groups based on **two factors**, with or without interaction.
4. **F-test in regression**:
   Tests whether the regression model as a whole is statistically significant.

Assumptions

- **Normal Distribution**: The data should follow a normal distribution.
- **Homogeneity of Variances**: The variances across groups should be equal.

Variance between groups measures the spread of group means around the grand mean.

- **Variance within groups** measures the spread of individual data points within each group.

Example

Suppose you have three groups of students with different teaching methods. You want to determine if there is a significant difference in their average scores.

Group	Mean Score	Variance
A	80	100
B	85	120
C	90	150

First, calculate the grand mean and then the variance between and within groups. If the calculated F-value exceeds the critical value from the F-distribution table, you reject the null hypothesis that all group means are equal.

Chi-Square Test

Purpose and Applications

The Chi-square test is a non-parametric test used to analyze categorical data and test relationships between nominal variables. It is commonly used for:

Types:

1. **Chi-Square Test of Independence**
    - Tests whether two categorical variables are independent of each other.
    - Example: Is gender independent of voting preference?
2. **Chi-Square Goodness-of-Fit Test**
    - Tests whether observed frequencies match expected frequencies.
    - Example: Are die rolls uniformly distributed?

Assumptions

- **Categorical Data**: The data must be categorical.
- **Expected Frequency**: The expected frequency in each category should be at least five.

Example

Suppose you want to determine if there is a relationship between gender and preference for a certain product. You collect data as follows:

Gender	Prefers Product	Does Not Prefer
Male	40	60
Female	30	70

Calculate the expected frequencies under the assumption of independence and then compute the Chi-square statistic. If the calculated value exceeds the critical value from the Chi-square distribution table, you reject the null hypothesis that the variables are independent.

Conclusion

Tests of significance are essential tools in statistical analysis, allowing researchers to make informed decisions about their hypotheses. Understanding the Z-test, t-test, F-test, and Chi-square test provides a solid foundation for conducting statistical inference in various fields.

Key Points Summary

- **Z-Test**: Used for large samples with known variance.
- **t-Test**: Used for small samples or unknown variance.

- **F-Test (ANOVA)**: Compares means across multiple groups.
- **Chi-Square Test**: Analyzes categorical data relationships.

Each test has specific assumptions and applications, and choosing the right test depends on the nature of the data and the research question.

Comparison Table

Test	Use	Sample Size	Distribution Assumed	Data Type	Key Features
Z-Test	Compare means or proportions	Large (n > 30)	Normal (Z-distribution)	Interval/Ratio	Used when population variance is known or large n
T-Test	Compare sample means (1 or 2 samples)	Small (n ≤ 30)	Student's t-distribution	Interval/Ratio	Population variance unknown; requires normality
F-Test	Compare variances; ANOVA	Any	F-distribution	Interval/Ratio	Right-tailed; used in regression and ANOVA
Chi-Square	Test for independence or goodness-of-fit	Any	Chi-square distribution	Categorical	Non-parametric; used for frequency/count data

Correlation and Regression

CORRELATION

Meaning:

- **Correlation** is a statistical measure that assesses the **degree and direction** of the relationship between two or more variables.
- The relationship indicates how one variable **changes with respect to** another. It's crucial to note that **correlation does not imply causality** — just because two variables are correlated, it doesn't mean that one causes the other.

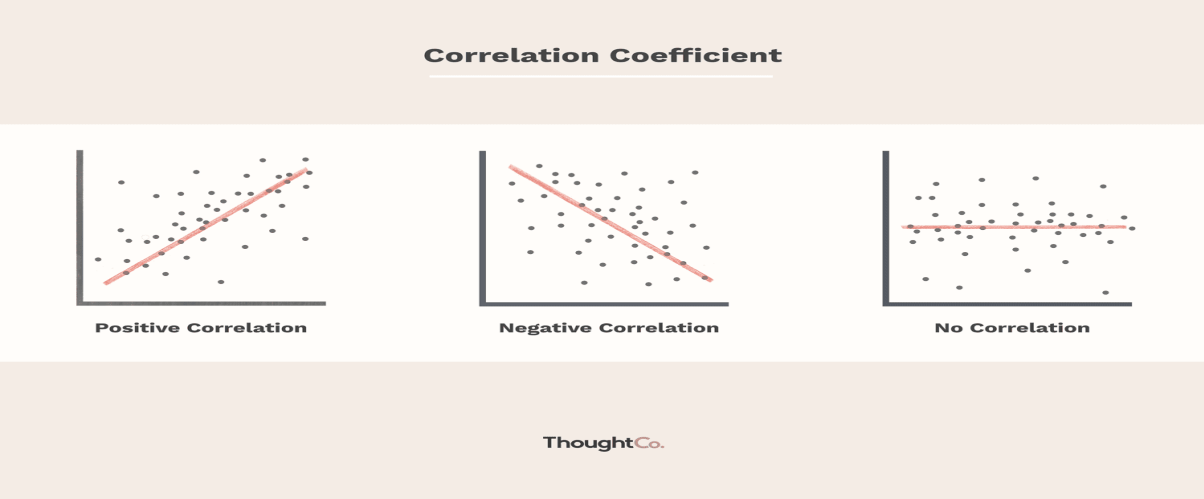
Example:

- A positive correlation could be between **education level** and **income**. As education increases, income tends to increase as well.
- A negative correlation could be between **gas prices** and **car sales**. As gas prices rise, car sales may decline, especially for gas-guzzling cars.

Types of Correlation:

1. Direction-based:

- **Positive Correlation**:
    - Both variables increase or decrease in the same direction.
    - Example: **Height** and **Weight** — as height increases, weight tends to increase as well.
- **Negative Correlation**:
    - One variable increases, and the other decreases.
    - Example: **Temperature** and **Coat Sales** — as the temperature rises, coat sales generally decrease.
- **Zero Correlation**:
    - No relationship between the variables.
    - Example: **Shoe size** and **IQ score** — no predictable relationship.



2. Based on the number of variables:

- **Simple Correlation**:
    - Involves two variables. It's the basic case of correlation, such as the relationship between income and expenditure.
- **Partial Correlation**:
    - Measures the relationship between two variables while controlling for the influence of one or more other variables.
    - Example: The correlation between exercise and health while controlling for diet.

- **Multiple Correlation**:
    - Involves more than two variables and assesses the relationship between one dependent variable and several independent variables.
    - Example: Studying how **income**, **education**, and **age** together affect **expenditure patterns**.

3. Nature of the Relationship:

- **Linear Correlation**:
    - A relationship where the change in one variable results in a proportional change in the other variable, and the relationship is represented by a straight line.
    - Example: **Price and quantity demanded** for a commodity.
- **Non-linear Correlation**:
    - The relationship between variables is not proportional and may involve curvatures or varying rates of change.
    - Example: **Growth of plants** over time, where growth accelerates initially and then slows down after reaching a certain limit.

Methods of Measuring Correlation:

1. **Graphical Methods**:
    - **Scatter Diagram**:
      A scatter plot visualizes the relationship between two variables. Points are plotted on the x-axis and y-axis, and the pattern of the points reveals the nature of the relationship.
2. **Algebraic Methods**:
    - **Karl Pearson's Correlation Coefficient** rr:
      This coefficient quantifies the degree of linear relationship between two variables.
      Formula: r lies between -1 and +1.
        - **Interpretation**:
            - r = +1: Perfect positive correlation (both variables move in the same direction).
            - r = -1: Perfect negative correlation (one variable increases, the other decreases).
            - r = 0: No correlation.

- The value of **r** helps in **predicting the strength** of the relationship.

Properties of the Correlation Coefficient:

- **Value Range**: The correlation coefficient rr always lies between **-1 and +1**.
- **Unitless**: It does not depend on the units of measurement of the variables.
- **Independent of Origin**: Shifting the origin of the variables does not change the correlation.
- **Symmetry**: The correlation between X and Y is the same as the correlation between Y and X.

REGRESSION

Meaning:

- **Regression** analysis helps to **predict** the value of one variable (dependent variable) based on the value of another (independent variable).
- It helps in understanding the **magnitude** and **direction** of the relationship and in making predictions based on past data.

Key Concepts:

- **Dependent Variable (Y)**: The variable you want to predict or explain.
- **Independent Variable (X)**: The variable used to make predictions about Y.

Example:

- In **sales forecasting**, the **sales** (Y) can be predicted based on **advertising spend** (X).

Uses of Regression Analysis:

1. It provides estimates of values of the dependent variables from values of independent variables.
2. It is used to obtain a measure of the error involved in using the regression line as a basis for estimation.
3. With the help of regression analysis, we can obtain a measure of degree of association or correlation that exists between the two variables.
4. It is highly valuable tool in economies and business research, since most of the problems of the economic analysis are based on cause and effect relationship.

Regression Coefficients:

- The **regression coefficient** bb represents the **slope** of the regression line, or the **change in Y** for a **unit change in X**.

Properties of Regression Coefficients:

- The **sign** of the regression coefficient (positive or negative) corresponds to the direction of the relationship between the variables.
- If one regression coefficient is **greater than 1**, the other will be **less than 1** due to the relationship with the correlation coefficient.
- **Regression coefficients** are independent of the **origin** but dependent on the **scale** of the variables.

Distinction between Correlation and Regression

Sl No	**Correlation**	**Regression**
1	It measures the degree and direction of relationship between the variables.	It measures the nature and extent of average relationship between two or more variables in terms of the original units of the data.
2	It is a relative measure showing association between the variables.	It is an absolute measure of relationship.
3	Correlation Coefficient is independent of change of both origin and scale.	Regression Coefficient is independent of change of origin but not scale.
4	Correlation Coefficient is independent of units of measurement.	Regression Coefficient is not independent of units of measurement.
5	Expression of the relationship between the variables ranges from –1 to +1.	Expression of the relationship between the variables may be in any of the forms like: $Y = a + bX$ $Y = a + bX + cX^2$
6	It is not a forecasting device.	It is a forecasting device which can be used to predict the value of dependent variable from the given value of independent variable.