Hypothesis Testing

Hypothesis testing is a statistical method used to make decisions or inferences about population parameters based on sample data. It typically involves:

1. Formulating two hypotheses: the **null hypothesis** (H_0) and the **alternative hypothesis** (H_1).
2. Selecting a significance level, often denoted by α (commonly 0.05 or 5%).
3. Choosing an appropriate test (e.g., t-test, z-test).
4. Calculating the test statistic and comparing it to the critical value or using the p-value.
5. Making a decision: rejecting or failing to reject H_0 based on the results.

Hypothesis Testing on a Dataset

Let's test whether there is a significant difference between the average weights of two groups (e.g., males and females). Suppose we're using a dataset from Kaggle with information on individuals' weights.

1. Define Hypotheses

- H0: The mean weight of males is equal to the mean weight of females (μmales=μfemales).
- H1: The mean weight of males is different from the mean weight of females (μmales≠μfemales).

2. Choose a Significance Level

- We choose α=0.05

3. Select the Test

- Since we are comparing the means of two independent groups, we use a **two-sample t-test**.

4. Perform the Test and Calculate p-value

Assume the dataset (weights.csv) has columns Gender (Male/Female) and Weight (numeric weight values).

See ***hypothesis-weights.py***

If p-value < 0.05: There is enough statistical evidence to reject the null hypothesis. We conclude that the average weights of males and females are significantly different.

If p-value ≥ 0.05: We do not have enough evidence to reject the null hypothesis, so we conclude that there is no significant difference in average weights.

Basic Test Types

Here's a breakdown of various hypothesis tests with examples and explanations for when each type is appropriate. We'll cover:

1. **t-tests** (One-sample, Two-sample, Paired)
2. **ANOVA (Analysis of Variance)**
3. **Chi-square tests**
4. **Correlation tests (Pearson, Spearman)**
5. **Regression analysis tests**

t-tests

A t test is a statistical hypothesis test that assesses sample means to draw conclusions about population means. Frequently, analysts use a t test to determine whether the population means for two groups are different. For example, it can determine whether the difference between the treatment and control group means is statistically significant.

The following are the standard t tests:

- One-sample: Compares a sample mean to a reference value.
- Two-sample: Compares two sample means.
- Paired: Compares the means of matched pairs, such as before and after scores.

ANOVA

ANOVA (Analysis of Variance) is a statistical method used to test if there are statistically significant differences between the means of three or more independent groups. It helps answer questions like, "Are the average scores of students in different classes significantly different?" or "Do three store locations have different average monthly sales?"

Key Concepts

- **Null Hypothesis (H0)**: All group means are equal.
- **Alternative Hypothesis (H1)** At least one group mean is different.
- **F-statistic**: The ratio of the variance between group means to the variance within the groups.
- **p-value**: If the p-value is less than the chosen significance level (usually 0.05), we reject the null hypothesis, suggesting a significant difference between at least two group means.

Example Scenario

Suppose we want to test if three different study methods lead to different average test scores among students.

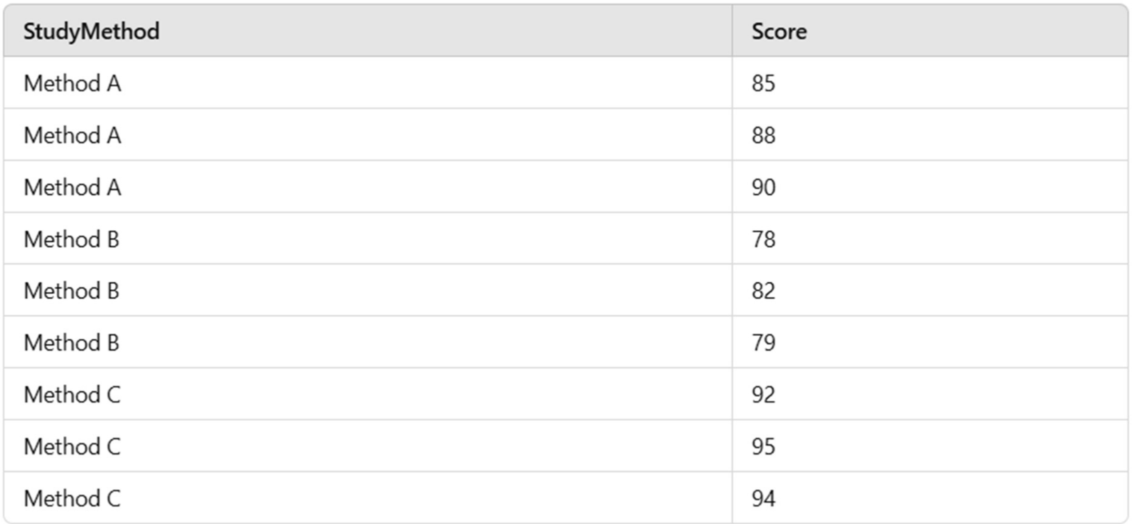
Hypotheses

- **Null Hypothesis (H0)**: The average test scores are the same across all study methods.
- **Alternative Hypothesis (H1)** At least one study method has a different average test score.

Dataset Example

We have a dataset study_methods.csv with two columns:

- StudyMethod: The study method used by each student (values: Method A, Method B, Method C).
- Score: The test score obtained by each student.



StudyMethod	Score
Method A	85
Method A	88
Method A	90
Method B	78
Method B	82
Method B	79
Method C	92
Method C	95
Method C	94

Steps for Conducting ANOVA

1. **Calculate Mean Scores**: Compute the mean score for each study method group.
2. **Calculate Variance Between and Within Groups**:
   - **Between-group variance**: How much the means of each group differ from the overall mean.
   - **Within-group variance**: How much scores vary within each group.
3. **Calculate the F-statistic**:

$$F = \frac{\text{Variance Between Groups}}{\text{Variance Within Groups}}$$

4. **Determine the p-value**: If the p-value is below 0.05, reject the null hypothesis.

Interpreting the Results

- **F-statistic**: A larger F-statistic indicates that the variability between group means is higher than the variability within groups.
- **p-value**: If the p-value is below 0.05, we reject the null hypothesis and conclude that at least one group's mean score differs significantly.

To illustrate ANOVA further, let's consider an example using the popular **Heart Disease Dataset**. This dataset is available on platforms like Kaggle and contains medical records used to predict the presence of heart disease. It includes various features, such as age, cholesterol levels, and types of chest pain.

Problem Statement

Suppose we want to analyze if there's a significant difference in the average cholesterol levels between people who experience different types of chest pain. This will help us understand if chest pain type could be associated with higher or lower cholesterol levels, potentially contributing to heart disease.

Dataset Overview

The Heart Disease Dataset includes the following columns (among others):

- **age**: Age of the patient
- **sex**: Gender (1 = male; 0 = female)
- **cp**: Chest pain type (categorical):
    - 0: Typical angina
    - 1: Atypical angina
    - 2: Non-anginal pain
    - 3: Asymptomatic
- **chol**: Serum cholesterol level in mg/dl

Objective

To use **ANOVA** to test if there's a statistically significant difference in the average cholesterol levels across the different chest pain types.

Hypotheses

- **Null Hypothesis (H0)**: The average cholesterol levels are the same across all chest pain types.
- **Alternative Hypothesis (H1)**: At least one chest pain type has a different average cholesterol level.

Steps to Perform ANOVA

1. **Prepare the Data**: Filter the data to include relevant columns (chest pain type and cholesterol levels).
2. **Calculate Means**: Calculate the mean cholesterol level for each chest pain type.
3. **Perform ANOVA**: Use the f_oneway function from the scipy.stats library to calculate the F-statistic and p-value.
4. **Interpret the Results**: Based on the p-value, determine if there's a significant difference in cholesterol levels between the groups.

See python code

Explanation of Results

- **F-statistic**: A higher F-statistic value indicates a greater variance between group means compared to within groups, suggesting that the groups' cholesterol levels differ.
- **p-value**: The p-value indicates the probability of observing the data if the null hypothesis were true.

If the p-value is less than 0.05, we reject the null hypothesis, suggesting that there is a statistically significant difference in average cholesterol levels among the different types of chest pain. This could indicate a relationship between chest pain type and cholesterol levels, which might warrant further medical investigation.

Example Interpretation

Let's say the output gives:

- **F-statistic**: 4.56
- **p-value**: 0.003

Since the p-value (0.003) is less than 0.05, we would **reject the null hypothesis** and conclude that there is a statistically significant difference in cholesterol levels among patients with different chest pain types. This might imply that cholesterol levels could be associated with certain chest pain types, providing valuable insights for diagnosing and understanding heart disease risks.

Chi-square Test

The Chi-square test is a statistical method used to determine if there is a significant association between two categorical variables. It assesses how expectations compare to actual observed data. The test is widely used in hypothesis testing, especially in studies involving frequency counts in contingency tables.

Types of Chi-square Tests

1. **Chi-square Test of Independence**: Tests whether two categorical variables are independent of each other.
2. **Chi-square Goodness of Fit Test**: Tests whether the distribution of a single categorical variable fits a specified distribution.

Example: Chi-square Test of Independence

Scenario

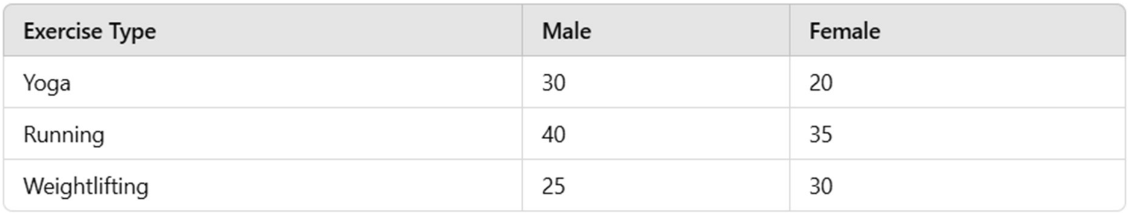
Suppose we want to examine whether there is an association between gender and preference for a type of exercise (e.g., yoga, running, weightlifting). We collect data from a sample of individuals regarding their gender and preferred exercise type.

Hypotheses

- **Null Hypothesis (H0)**: There is no association between gender and exercise preference (the variables are independent).
- **Alternative Hypothesis (H1)**: There is an association between gender and exercise preference (the variables are not independent).

Dataset Overview

We can create a hypothetical dataset for this scenario that looks like the following contingency table:



Exercise Type	Male	Female
Yoga	30	20
Running	40	35
Weightlifting	25	30

See python code

Performing the Chi-square Test in Python

1. **Organize the Data**: Create a contingency table.
2. **Conduct the Chi-square Test**: Use the chi2_contingency function from the scipy.stats library.

Explanation of the Code

1. **Create the Contingency Table**: We set up our data in a structured format, representing counts of preferences by gender.
2. **Chi-square Test**: The chi2_contingency function calculates the Chi-square statistic, p-value, degrees of freedom, and the expected frequencies based on the observed counts.
3. **Interpretation**: We compare the p-value to our significance level (usually 0.05) to determine whether to reject or fail to reject the null hypothesis.

Interpreting the Results

- **Chi-square Statistic**: A higher value indicates a greater difference between observed and expected frequencies.
- **p-value**: If the p-value is less than 0.05, we reject the null hypothesis, indicating that there is a significant association between gender and exercise preference.

Example Output

Let's say the output gives:

- **Chi-square Statistic**: 4.56
- **p-value**: 0.033

Since the p-value (0.033) is less than 0.05, we **reject the null hypothesis** and conclude that there is a significant association between gender and exercise preference. This finding suggests that exercise preferences may vary between males and females.

Chi-square for heart dataset

Let’s enhance the explanation of the Chi-square test using the **Heart Disease Dataset**. This dataset can help us analyze the relationship between different categorical variables, such as the presence of heart disease and other factors like chest pain type or smoking status.

Scenario

We want to explore whether there is an association between the type of chest pain (categorical variable) and the presence of heart disease (another categorical variable). This can provide insights into how chest pain might be related to heart disease risk.

Hypotheses

- **Null Hypothesis (H0)**: There is no association between chest pain type and the presence of heart disease (the variables are independent).
- **Alternative Hypothesis (H1)**: There is an association between chest pain type and the presence of heart disease (the variables are not independent).

Dataset Overview

In the **Heart Disease Dataset**, relevant columns might include:

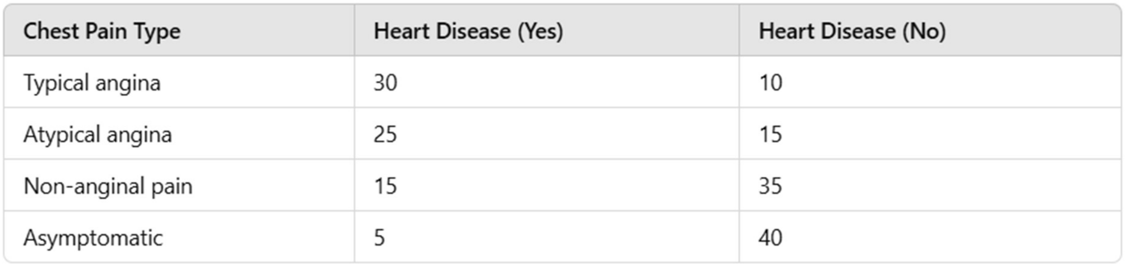
- **cp**: Chest pain type (0, 1, 2, 3).
    - 0: Typical angina
    - 1: Atypical angina
    - 2: Non-anginal pain
    - 3: Asymptomatic
- **target**: Presence of heart disease (1 = presence, 0 = absence).

Preparing the Data

We'll create a contingency table that counts the number of cases for each combination of chest pain type and heart disease presence.

Example Contingency Table

Suppose our analysis results in the following counts:



Chest Pain Type	Heart Disease (Yes)	Heart Disease (No)
Typical angina	30	10
Atypical angina	25	15
Non-anginal pain	15	35
Asymptomatic	5	40

See python code

Interpreting the Results

- **Chi-square Statistic**: A higher Chi-square statistic value indicates that there is a greater difference between the observed and expected counts, suggesting a possible association between the variables.
- **p-value**: If the p-value is less than 0.05, we reject the null hypothesis, indicating that there is a significant association between chest pain type and heart disease presence.

Example Output

Let's assume the output is as follows:

- **Chi-square Statistic**: 12.34
- **p-value**: 0.002

Since the p-value (0.002) is less than 0.05, we **reject the null hypothesis** and conclude that there is a significant association between chest pain type and the presence of heart disease. This result suggests that certain types of chest pain may be linked to a higher risk of heart disease.

Conclusion

The Chi-square test is an essential tool for analyzing categorical data, helping identify relationships between variables. In the context of the Heart Disease Dataset, it provides valuable insights into how chest pain type relates to heart disease risk. These findings can inform medical professionals and researchers in diagnosing and understanding heart disease and its associated factors.

Correlation Tests

Pearson Correlation Test

- **Scenario**: Measures the linear relationship between two continuous variables.

Spearman Correlation Test

- **Scenario**: Measures monotonic relationships (not necessarily linear) between two continuous or ordinal variables.

Regression Analysis Tests

Regression tests analyze relationships between one dependent variable and one or more independent variables.