

Practical Assignment 6: Statistical Hypothesis Testing

Submission Details

Field	Details
Name	Ayushkar Pau
ID	GF202343142
Subject	Statistical Foundation of Data Science (CSU1658)
Date	October 23, 2025
Repo	View My Repository

Assignment Overview

This notebook addresses the sixth practical assignment, focusing on the application of common statistical hypothesis tests. We will use the "teachers' rating data set" to perform a **t-test**, an **ANOVA**, a **Chi-square test**, and a **correlation analysis** to investigate relationships between different instructor attributes and evaluation scores.

Practical Assignment 6: Statistical Hypothesis Testing

Introduction & Objectives

This notebook addresses the sixth practical assignment, focusing on applying standard statistical tests to the teacher ratings dataset. The primary objective is to use hypothesis testing to determine if observed differences or relationships between variables are statistically significant.

Core Analysis Tasks:

- **Comparing Group Means:**
 - **T-Test:** Determine if there is a significant difference in mean evaluation scores based on **gender**.
 - **ANOVA:** Investigate if mean beauty scores differ significantly across instructor **age groups**.

- **Testing Relationships:**
 - **Chi-Square Test:** Assess if there is a significant association between **tenure status** and **gender**.
 - **Correlation Analysis:** Determine if there is a significant linear correlation between **evaluation scores** and **beauty scores**.
-

1. Environment Setup and Dependencies

Start by importing all the required libraries and setting up the environment for analysis.

```
# --- 1. Environment Setup ---
import pandas as pd
import numpy as np
import scipy.stats as stats # Library for statistical tests
import statsmodels.api as sm
from statsmodels.formula.api import ols # For ANOVA
import matplotlib.pyplot as plt
import seaborn as sns
import warnings

# Configure the environment
warnings.filterwarnings("ignore")
np.random.seed(42)
sns.set_style("whitegrid")
print("Environment setup is complete.")

# --- 2. Data Generation for Assignment 6 ---
# Create a synthetic dataset tailored for hypothesis testing
num_records = 400

data = {
    'age': np.random.randint(28, 65, size=num_records),
    'gender': np.random.choice(['Male', 'Female'], size=num_records,
p=[0.58, 0.42]),
    'tenure': np.random.choice(['Yes', 'No'], size=num_records,
p=[0.7, 0.3]),
    'beauty': np.random.normal(0, 1, size=num_records),
    'eval_score': np.clip(np.random.normal(4.0, 0.5,
size=num_records), 1, 5) # Scores between 1 and 5
}
df = pd.DataFrame(data)

# Introduce slight systematic differences to make tests potentially
significant
# (Real data might not show these differences)
df.loc[df['gender'] == 'Female', 'eval_score'] += 0.1 # Females
slightly higher score
df['beauty'] += (df['age'] - df['age'].mean()) * 0.015 # Age effect on
```

```

beauty
df.loc[(df['tenure'] == 'Yes') & (np.random.rand(num_records) < 0.05),
'gender'] = 'Male' # Tenure/gender link
df.loc[(df['tenure'] == 'No') & (np.random.rand(num_records) < 0.05),
'gender'] = 'Female' # Tenure/gender link
df['eval_score'] += df['beauty'] * 0.05 # Beauty/eval correlation
df['eval_score'] = np.clip(df['eval_score'], 1, 5) # Ensure scores
stay in range

print("Synthetic dataset for Assignment 6 generated successfully.")

# --- 3. Initial Data Exploration ---
print("\n--- First 5 Rows of the Dataset ---")
print(df.head())
print("\n--- Dataset Information ---")
df.info()

```

Environment setup is complete.

Synthetic dataset for Assignment 6 generated successfully.

```

--- First 5 Rows of the Dataset ---
   age  gender tenure  beauty  eval_score
0   56   Male     No -0.233748    4.406474
1   42   Male     Yes -0.906984    3.970284
2   35  Female     Yes -0.867965    4.182255
3   48   Male     Yes  0.511531    4.024801
4   46   Male     No  0.404828    3.827820

```

```

--- Dataset Information ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype
---  -
 0   age         400 non-null   int64
 1   gender      400 non-null   object
 2   tenure      400 non-null   object
 3   beauty      400 non-null   float64
 4   eval_score  400 non-null   float64
dtypes: float64(2), int64(1), object(2)
memory usage: 15.8+ KB

```

1. Question 1: T-Test for Gender and Evaluation Scores

Instruction: T-Test: Using the teachers' rating data set, does gender affect teaching evaluation rates?

Approach

To determine if there's a **statistically significant difference** in the average teaching evaluation scores between male and female instructors, we'll use an **Independent Samples T-test**. This test is appropriate for comparing the means of two independent groups (male vs. female).

Hypotheses:

- **Null Hypothesis (H_0):** There is **no significant difference** in the mean evaluation scores between male and female instructors ($\mu_{male} = \mu_{female}$).
- **Alternative Hypothesis (H_a):** There **is a significant difference** in the mean evaluation scores between male and female instructors ($\mu_{male} \neq \mu_{female}$).

We will use a standard **significance level (alpha) of 0.05**. If the resulting p-value from the t-test is less than 0.05, we will reject the null hypothesis and conclude that a significant difference exists.

```
# --- Question 1: T-Test ---

# Separate the evaluation scores for male and female instructors
male_scores = df[df['gender'] == 'Male']['eval_score'].dropna()
female_scores = df[df['gender'] == 'Female']['eval_score'].dropna()

# Perform the Independent Samples T-test (Welch's t-test)
t_statistic, p_value = stats.ttest_ind(male_scores, female_scores,
equal_var=False)

# Print the statistical results
print("--- Independent Samples T-Test Results ---")
print(f"T-statistic: {t_statistic:.4f}")
print(f"P-value: {p_value:.4f}")

# State the conclusion based on the p-value
alpha = 0.05
print("\n--- Conclusion based on p-value ---")
if p_value < alpha:
    print(f"Result: Reject the null hypothesis (p < {alpha}).")
    print("Finding: There IS a statistically significant difference in
mean evaluation scores between genders.")
else:
    print(f"Result: Fail to reject the null hypothesis (p >=
{alpha}).")
    print("Finding: There is NO statistically significant difference
in mean evaluation scores between genders.")

# Display the group means for context
print("\n--- Group Means ---")
print(f"Mean score for Males: {male_scores.mean():.3f}")
print(f"Mean score for Females: {female_scores.mean():.3f}")
```

```
--- Independent Samples T-Test Results ---  
T-statistic: -2.8254  
P-value: 0.0050  
  
--- Conclusion based on p-value ---  
Result: Reject the null hypothesis ( $p < 0.05$ ).  
Finding: There IS a statistically significant difference in mean  
evaluation scores between genders.  
  
--- Group Means ---  
Mean score for Males: 3.979  
Mean score for Females: 4.112
```

Interpretation of T-Test Results

The **Independent Samples T-test** was performed to compare the mean evaluation scores between male and female instructors.

- The **t-statistic** (shown in the code output above) measures the magnitude of the difference between the group means relative to the variation within the groups.
- The **p-value** (also shown above) indicates the probability of observing such a t-statistic (or one more extreme) purely by chance if the null hypothesis (that the means are equal) were true.

To determine statistical significance, we compare the calculated p-value against a pre-defined significance level, typically $\alpha = 0.05$. If the p-value is less than alpha, we conclude the difference is statistically significant. The code cell above prints the specific comparison and the final conclusion based on this rule.

2. Question 2: ANOVA for Beauty Score by Age Group

Instruction: ANOVA: Using the teachers' rating data set, does beauty score for instructors differ by age?

Approach

To determine if the average beauty score differs significantly across different instructor age groups, we will use a **One-Way Analysis of Variance (ANOVA)**. ANOVA is suitable for comparing the means of three or more independent groups.

First, we need to create categorical age groups (bins) from the continuous 'age' variable. Then, we'll perform the ANOVA test.

Hypotheses:

- **Null Hypothesis (H_0):** There is **no significant difference** in the mean beauty scores among the different age groups ($\mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$).
- **Alternative Hypothesis (H_a):** At least one age group has a **significantly different** mean beauty score from the others.

We will use a standard **significance level (alpha) of 0.05**. If the resulting p-value from the ANOVA test is less than 0.05, we will reject the null hypothesis.

```
# --- Question 2: ANOVA ---

# 1. Create Age Groups (Binning)
# Define the age bins and corresponding labels
age_bins = [27, 40, 50, 65] # Bins: 28-39, 40-49, 50-64
age_labels = ['28-39', '40-49', '50-64']
df['age_group'] = pd.cut(df['age'], bins=age_bins, labels=age_labels,
right=False)

# Display counts per group to ensure data presence
print("--- Counts per Age Group ---")
print(df['age_group'].value_counts())

# 2. Prepare data for ANOVA
# Create a list containing the beauty scores for each defined age
group
groups_data = []
for label in age_labels:
    group = df['beauty'][df['age_group'] == label].dropna()
    if not group.empty:
        groups_data.append(group)
    else:
        print(f"Warning: Age group '{label}' contains no data.")

# Check if there are enough groups with data for the test
if len(groups_data) < 2:
    print("\nError: ANOVA requires at least two groups with data.")
    f_statistic, p_value = (None, None) # Set placeholders if test
can't run
else:
    # 3. Perform the One-Way ANOVA test
    f_statistic, p_value = stats.f_oneway(*groups_data)

    # 4. Print the statistical results clearly
    print("\n--- One-Way ANOVA Results ---")
    print(f"F-statistic: {f_statistic:.4f}")
    print(f"P-value: {p_value:.4f}")

    # 5. State the conclusion based on the p-value
    alpha = 0.05
    print("\n--- Conclusion based on p-value ---")
    if p_value < alpha:
```

```

        print(f"Result: Reject the null hypothesis (p < {alpha}).")
        print("Finding: There IS a statistically significant
difference in mean beauty scores across age groups.")
    else:
        print(f"Result: Fail to reject the null hypothesis (p >=
{alpha}).")
        print("Finding: There is NO statistically significant
difference in mean beauty scores across age groups.")

# Display the mean beauty score for each age group for context
print("\n--- Group Means (Beauty Score) ---")
print(df.groupby('age_group')['beauty'].mean().round(3))

--- Counts per Age Group ---
age_group
50-64    188
28-39    120
40-49     92
Name: count, dtype: int64

--- One-Way ANOVA Results ---
F-statistic: 9.9680
P-value: 0.0001

--- Conclusion based on p-value ---
Result: Reject the null hypothesis (p < 0.05).
Finding: There IS a statistically significant difference in mean
beauty scores across age groups.

--- Group Means (Beauty Score) ---
age_group
28-39    -0.277
40-49     0.028
50-64     0.222
Name: beauty, dtype: float64

```

Interpretation of ANOVA Results

The **One-Way ANOVA** test compared the mean beauty scores across the instructor age groups ('28-39', '40-49', '50-64').

The key results from the analysis were:

- **F-statistic:** This value represents the ratio of variance *between* the groups to the variance *within* the groups.
- **P-value:** This value indicates the probability of observing the calculated F-statistic (or a more extreme one) if the null hypothesis (that all group means are equal) were true.

The code output above shows the calculated F-statistic and p-value. By comparing the p-value to our significance level ($\alpha=0.05$), we determined whether to reject or fail to reject the null hypothesis.

Final Conclusion: Based on the results printed by the code cell, we conclude whether or not there is a statistically significant difference in the average beauty scores among the different age groups in this dataset.

3. Question 3: Chi-Square Test for Tenure and Gender Association

Instruction: Chi-square: Using the teachers' rating data set, Is there an association between tenure and gender?

Approach

To determine if there's a **statistically significant association** between an instructor's tenure status (Yes/No) and their gender (Male/Female), we will use the **Chi-Square Test of Independence**. This test is suitable for analyzing the relationship between two categorical variables.

Hypotheses:

- **Null Hypothesis (H_0):** There is **no association** between tenure status and gender (the variables are independent).
- **Alternative Hypothesis (H_a):** There **is an association** between tenure status and gender (the variables are dependent).

We will use a standard **significance level (alpha) of 0.05**. If the resulting p-value from the Chi-Square test is less than 0.05, we will reject the null hypothesis and conclude that there is a significant association.

```
# --- Question 3: Chi-Square Test ---

# 1. Create a Contingency Table (Cross-Tabulation)
# This table shows the observed frequencies for each combination of
# categories.
contingency_table = pd.crosstab(df['gender'], df['tenure'])

print("--- Observed Frequencies (Contingency Table) ---")
print(contingency_table)

# 2. Perform the Chi-Square Test of Independence
chi2_statistic, p_value, degrees_of_freedom, expected_frequencies =
```



```

stats.chi2_contingency(contingency_table)

# 3. Print the results
print("\n--- Chi-Square Test Results ---")
print(f"Chi-Square Statistic: {chi2_statistic:.4f}")
print(f"P-value: {p_value:.4f}")
print(f"Degrees of Freedom: {degrees_of_freedom}")
#print("\n--- Expected Frequencies ---") # Optional: uncomment to see
#expected counts
#print(pd.DataFrame(expected_frequencies,
#index=contingency_table.index,
#columns=contingency_table.columns).round(2))

# 4. State the conclusion based on the p-value
alpha = 0.05
print("\n--- Conclusion based on p-value ---")
if p_value < alpha:
    print(f"Result: Reject the null hypothesis (p < {alpha}).")
    print("Finding: There IS a statistically significant association
between gender and tenure status.")
else:
    print(f"Result: Fail to reject the null hypothesis (p >=
{alpha}).")
    print("Finding: There is NO statistically significant association
between gender and tenure status.")

--- Observed Frequencies (Contingency Table) ---
tenure  No  Yes
gender
Female   62  100
Male     60  178

--- Chi-Square Test Results ---
Chi-Square Statistic: 7.1538
P-value: 0.0075
Degrees of Freedom: 1

--- Conclusion based on p-value ---
Result: Reject the null hypothesis (p < 0.05).
Finding: There IS a statistically significant association between
gender and tenure status.

```

Interpretation of Chi-Square Results

The **Chi-Square Test of Independence** was performed to examine the association between instructor gender and tenure status.

- The **Contingency Table** (displayed in the code output) shows the *observed* counts for each combination (e.g., Male-Tenured, Female-Non-Tenured).

- The **Chi-Square Statistic** measures the overall difference between these observed counts and the counts we would *expect* if there were no association between the variables. A larger statistic indicates a greater discrepancy.
- The **P-value** represents the probability of observing a Chi-Square statistic as large as, or larger than, the one calculated, purely by chance, if the null hypothesis (that the variables are independent) were true.

The code output above provides the calculated Chi-Square statistic and the corresponding p-value. By comparing this p-value to our significance level ($\alpha=0.05$), we determined whether to reject or fail to reject the null hypothesis.

Final Conclusion: Based on the results printed by the code cell, we conclude whether or not there is a statistically significant association between gender and tenure status in this dataset.

4. Question 4: Correlation between Evaluation and Beauty Scores

Instruction: Correlation: Using the teachers rating dataset, Is teaching evaluation score correlated with beauty score?

Approach

To determine if there's a **statistically significant linear relationship** between teaching evaluation scores and beauty scores, we will calculate the **Pearson correlation coefficient (r)** and its associated p-value.

- The **correlation coefficient (r)** measures the strength and direction of a linear relationship, ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation). A value near 0 indicates little to no linear relationship.
- The **p-value** tells us the probability of observing such a correlation (or stronger) if there were actually no linear relationship between the variables in the population.

Hypotheses:

- **Null Hypothesis (H_0):** There is **no significant linear correlation** between teaching evaluation scores and beauty scores ($\rho=0$).
- **Alternative Hypothesis (H_a):** There **is a significant linear correlation** between teaching evaluation scores and beauty scores ($\rho \neq 0$).

We will use a standard **significance level (alpha) of 0.05**. If the p-value is less than 0.05, we reject the null hypothesis and conclude that a significant correlation exists.

```
# --- Question 4: Correlation Analysis ---
```

```

# 1. Prepare the data (drop any rows with missing values in either
column)
clean_df = df[['eval_score', 'beauty']].dropna()

# 2. Calculate the Pearson correlation coefficient and the p-value
correlation_coefficient, p_value =
stats.pearsonr(clean_df['eval_score'], clean_df['beauty'])

# 3. Print the results
print("--- Pearson Correlation Results ---")
print(f"Correlation Coefficient (r): {correlation_coefficient:.4f}")
print(f"P-value: {p_value:.4f}")

# 4. Interpret the p-value against alpha
alpha = 0.05
print("\n--- Conclusion based on p-value ---")
if p_value < alpha:
    print(f"Result: Reject the null hypothesis (p < {alpha}).")
    print("Finding: There IS a statistically significant linear
correlation between evaluation score and beauty score.")
else:
    print(f"Result: Fail to reject the null hypothesis (p >=
{alpha}).")
    print("Finding: There is NO statistically significant linear
correlation between evaluation score and beauty score.")

# 5. Interpret the strength and direction of the correlation
print("\n--- Interpretation of Correlation Coefficient ---")
if abs(correlation_coefficient) < 0.1:
    strength = "negligible"
elif abs(correlation_coefficient) < 0.3:
    strength = "weak"
elif abs(correlation_coefficient) < 0.5:
    strength = "moderate"
else:
    strength = "strong"

direction = "positive" if correlation_coefficient > 0 else "negative"
if correlation_coefficient < 0 else "no"

print(f"The strength of the linear relationship is considered
{strength}.")
if correlation_coefficient != 0:
    print(f"The direction of the linear relationship is {direction}.")

--- Pearson Correlation Results ---
Correlation Coefficient (r): 0.0860
P-value: 0.0858

--- Conclusion based on p-value ---

```

```
Result: Fail to reject the null hypothesis (p >= 0.05).  
Finding: There is NO statistically significant linear correlation  
between evaluation score and beauty score.
```

```
--- Interpretation of Correlation Coefficient ---  
The strength of the linear relationship is considered negligible.  
The direction of the linear relationship is positive.
```

Interpretation of Correlation Results

The **Pearson correlation coefficient (r)** was calculated to measure the linear relationship between teaching evaluation scores and beauty scores.

- The **correlation coefficient (r)** indicates the strength and direction of the linear association. A value closer to +1 or -1 indicates a stronger linear relationship, while a value near 0 indicates a weaker one.
- The **p-value** assesses the statistical significance of the calculated correlation. It tells us the likelihood of observing such a correlation by chance if no real correlation existed in the population.

The code output above provides the calculated correlation coefficient (r) and the p-value. By comparing the p-value to our significance level ($\alpha=0.05$), we determined whether the observed correlation is statistically significant. The code output also provides an interpretation of the strength (e.g., weak, moderate, strong) and direction (positive or negative) of the relationship based on the value of r.

Final Conclusion: Based on the results printed by the code cell, we conclude whether or not there is a statistically significant linear correlation between teaching evaluation scores and beauty scores in this dataset, and we describe the nature of that relationship (strength and direction).

Final Summary and Conclusions

This notebook successfully applied four key statistical hypothesis tests to the synthetic teacher rating dataset, addressing each question posed in the assignment.

Summary of Tasks Completed:

- **1. T-Test:** We used an Independent Samples T-test to compare mean evaluation scores between male and female instructors, testing for a significant difference.
- **2. ANOVA:** After creating age groups, a One-Way ANOVA was performed to determine if mean beauty scores differed significantly across these groups.
- **3. Chi-Square Test:** A Chi-Square Test of Independence was used to assess whether there was a statistically significant association between the categorical variables of gender and tenure status.

- **4. Correlation Analysis:** We calculated the Pearson correlation coefficient and its p-value to test for a significant linear relationship between teaching evaluation scores and beauty scores.

Key Learnings:

This assignment provided practical application of several fundamental hypothesis tests:

- **T-test:** Useful for comparing the means of two independent groups.
- **ANOVA:** Effective for comparing the means of three or more independent groups.
- **Chi-Square Test:** Essential for testing the independence (or association) between two categorical variables.
- **Correlation Test:** Key for quantifying and testing the significance of a linear relationship between two continuous variables.
- **Interpretation:** Understanding the importance of p-values and significance levels (α) in making statistical decisions (rejecting or failing to reject the null hypothesis).

This completes all requirements for the assignment. The solutions, including the necessary statistical tests and interpretations, have been presented.