Practical Assignment 5: Probability and Hypothesis Testing

Submission Details

Field	Details
Name	Ayushkar Pau
ID	GF202343142
Subject	Statistical Foundation of Data Science (CSU1658)
Date	October 13, 2025

Assignment Overview

This notebook addresses the fifth practical assignment, focusing on two key areas of statistics. First, we will calculate empirical probabilities from the "Teacher Rating Dataset" to determine the likelihood of receiving certain evaluation scores. Second, we will formally state the null and alternative hypotheses for a two-tailed test comparing the performance of professional and regional basketball players.

1. Environment Setup and Dependencies

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

In [1]:
```
# --- 1. Environment Setup ---
import pandas as pd
import numpy as np
import warnings

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

# --- 2. Data Generation for Assignment 5 ---
# Create a synthetic "ratings" dataset
num_records = 500

data = {
    'eval_score': np.clip(np.random.normal(4.0, 0.5, size=num_records), 1, 5) # Score
}
df = pd.DataFrame(data)

print("Synthetic ratings dataset 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 ratings dataset generated successfully.

--- First 5 Rows of the Dataset ---
   eval_score
0    4.248357
1    3.930868
2    4.323844
3    4.761515
4    3.882923

--- Dataset Information ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 1 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   eval_score  500 non-null    float64
dtypes: float64(1)
memory usage: 4.0 KB
```

---

1. Question 1: Probability of Score > 4.5

> **Instruction**: Using the teachers' rating dataset, what Is the probability of receiving an evaluation score of greater than 4.5

Approach

To find the empirical probability, we will use the classic probability formula:

$$\text{P(Event)} = \frac{\text{Number of Favorable Outcomes}}{\text{Total Number of Outcomes}}$$

In our case, a "favorable outcome" is any evaluation score greater than 4.5. The "total number of outcomes" is the total number of records in our dataset.

In [4]:
```python
# --- Question 1: Probability of Score > 4.5 ---

# 1. Count the number of scores greater than 4.5
favorable_outcomes = df[df['eval_score'] > 4.5].shape[0]

# 2. Count the total number of records
total_outcomes = df.shape[0]

# 3. Calculate the probability
probability = favorable_outcomes / total_outcomes
percentage = probability * 100 # For easier printing

# 4. Print the dynamic result and interpretation
print(f"The probability of receiving an evaluation score greater than 4.5 is {probabi
print(f"\nThis means that based on our synthetic dataset, there is approximately a {p
```

```
The probability of receiving an evaluation score greater than 4.5 is 0.1420, which is equivalent to 14.20%.

This means that based on our synthetic dataset, there is approximately a 14.20% chance that any randomly selected teacher evaluation will have a score above 4.5.
```

Conclusion

The empirical probability was calculated by dividing the count of favorable outcomes (scores > 4.5) by the total number of observations.

The resulting low probability indicates that evaluation scores in the highest range are relatively rare in this dataset, making up only a small fraction of the total evaluations.

---

2. Question 2: Probability of Score between 3.5 and 4.2

> **Instruction**: Using the teachers' rating dataset, what Is the probability of receiving an evaluation score greater than 3.5 and less than 4.2

Approach

Similar to the first question, we will calculate the empirical probability. However, this time our "favorable outcome" has two conditions: the score must be **greater than 3.5** AND **less than 4.2**. We will filter our dataset to count the number of scores that meet both conditions and then divide by the total number of records.

In [5]:
```
# --- Question 2: Probability of Score between 3.5 and 4.2 ---

# 1. Count the number of scores that meet both conditions
favorable_outcomes = df[(df['eval_score'] > 3.5) & (df['eval_score'] < 4.2)].shape[0]

# 2. Count the total number of records
total_outcomes = df.shape[0]

# 3. Calculate the probability
probability = favorable_outcomes / total_outcomes
percentage = probability * 100

# 4. Print the dynamic result and interpretation
print(f"The probability of a score between 3.5 and 4.2 is {probability:.4f}, which is
print(f"\nThis means there is a {percentage:.2f}% chance that a randomly selected eva
```

```
The probability of a score between 3.5 and 4.2 is 0.5160, which is equivalent to 51.6
0%.

This means there is a 51.60% chance that a randomly selected evaluation will have a sc
ore in this range.
```

Conclusion

The probability was calculated by counting the number of evaluation scores that fell within the specified range (3.5 to 4.2) and dividing by the total number of observations.

This value represents the likelihood that a teacher's evaluation score falls within the central part of the rating distribution.

---

3. Question 3: Stating the Null Hypothesis

> **Instruction**: Using the two-talled test from a normal distribution... A professional basketball team wants to compare its performance with that of players in a regional

> league... The pro coach would like to know whether his professional team scores on average are different from that of the regional players. State the null hypothesis.

Approach

This is a conceptual question that requires us to frame the problem for a statistical test. We need to formally state the **Null Hypothesis (H_0)** and the **Alternative Hypothesis (H_a).**

- The **Null Hypothesis (H_0)** always represents the status quo or the position of "no difference." It's the statement we are trying to find evidence against.
- The **Alternative Hypothesis (H_a)** represents the new claim or the position that there *is* a difference.

Stated Hypotheses

Based on the scenario, the hypotheses for the two-tailed test are:

Null Hypothesis (H_0):

> The mean score of the regional players is **not different** from the historic mean of the professional players.
>
> Mathematically: $H_0 : \mu = 12$

Alternative Hypothesis (H_a):

> The mean score of the regional players **is different** from the historic mean of the professional players.
>
> Mathematically: $H_a : \mu \neq 12$

Interpretation

By setting up these hypotheses, we have framed the problem for a **two-tailed test**. The goal of such a test would be to analyze the sample data (the 10.7 points from 36 regional players) to determine if there is enough statistical evidence to **reject the null hypothesis**. If we reject it, we can conclude that the regional players' average score is statistically different from the professional players' historic average of 12.

Final Summary and Conclusions

This notebook successfully addressed all three problems for the fifth practical assignment, focusing on the core statistical concepts of empirical probability and hypothesis testing.

Summary of Tasks Completed:

- **1. Probability (Single Condition):** We calculated the empirical probability of a teacher receiving an evaluation score greater than 4.5 by filtering the dataset for favorable outcomes and dividing by the total number of observations.
- **2. Probability (Compound Condition):** We extended the probability calculation to a specific range, finding the likelihood of an evaluation score being greater than 3.5 *and* less than 4.2.

- **3. Hypothesis Testing:** For a given scenario comparing two basketball teams, we formally stated the **Null Hypothesis (H_0)** and the **Alternative Hypothesis (H_a)** for a two-tailed test, framing the problem for statistical analysis.

Key Learnings:

This assignment provided a solid foundation in two critical areas of statistics:

- **Empirical Probability:** Understanding how to calculate the real-world probability of an event by analyzing a sample dataset.
- **Hypothesis Framing:** Learning the crucial first step of any statistical test, which is to correctly define the null hypothesis (the statement of no difference) and the alternative hypothesis (the claim to be tested).

This completes all requirements for the assignment.