

Statistics Basics| Assignment

Q1: What is the difference between descriptive statistics and inferential statistics? Explain with examples.

Ans:

Descriptive statistics: Descriptive statistics focus on summarizing and describing the features of a specific dataset. It tells you what is happening in the data you have currently.

- **Example:** Calculating the average score of students in a single classroom to understand how that specific class performed.

Inferential statistics: Inferential statistics involve taking a sample from a larger population and making predictions or "inferences" about that population based on the sample.

- **Example:** Surveying 100 voters to predict the outcome of a national election involving millions of people.

Q2: What is sampling in statistics? Explain the differences between random and stratified sampling.

Ans:

Sampling is the process of selecting a subset of individuals from a statistical population to estimate characteristics of the whole group.

- **Random Sampling:** Every member of the population has an equal chance of being selected. It is like pulling names out of a hat.
- **Stratified Sampling:** The population is divided into subgroups (strata) based on shared characteristics (e.g., age or gender). A random sample is then taken from each subgroup to ensure they are all represented.

Q3: Define mean, median, and mode. Explain why these measures of central tendency are important.

Ans:

These are **measures of central tendency** used to find the "center" of a data distribution.

- **Mean:** The mathematical average, calculated by adding all values and dividing by the total count.
- **Median:** The middle value when the data is sorted in order.
- **Mode:** The value that appears most frequently in the dataset.

Importance:

- Describe central tendency
- Help understand data distribution
- Used in business, economics, and research

Q4: Explain skewness and kurtosis. What does a positive skew imply about the data?

Ans:

- **Skewness:** Measures the asymmetry of the probability distribution.
- **Kurtosis:** Measures the "tailedness" or the peakedness of the distribution.

Positive Skew: This implies that the tail on the right side of the distribution is longer or fatter than the left side. In a positively skewed dataset, the **Mean** is typically greater than the **Median**.

Q5: Implement a Python program to compute the mean, median, and mode of a given list of numbers.

numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]

Ans:

```
import statistics as stats

numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]

print("Mean:", stats.mean(numbers))
print("Median:", stats.median(numbers))
print("Mode:", stats.mode(numbers))
```

Output:

Mean: 19.6

Median: 19

Mode: 12

Q6: Compute the covariance and correlation coefficient between the following two datasets provided as lists in Python:

list_x = [10, 20, 30, 40, 50]

list_y = [15, 25, 35, 45, 60]

Ans:

```
import numpy as np
```

```
list_x = [10, 20, 30, 40, 50]
list_y = [15, 25, 35, 45, 60]

cov = np.cov(list_x, list_y)[0][1]
corr = np.corrcoef(list_x, list_y)[0][1]

print("Covariance:", cov)
print("Correlation:", corr)
```

Output:

Covariance: 275.0

Correlation: 0.996

Q7: Write a Python script to draw a boxplot for the following numeric list and identify its outliers. Explain the result:

data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]

Ans:

```
import matplotlib.pyplot as plt

data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]

plt.boxplot(data)
plt.title("Boxplot of Data")
plt.show()
```

Output:

The value 35 is likely identified as an outlier. In a boxplot, outliers are values that fall outside 1.5 times the Interquartile Range (IQR) above the third quartile or below the first quartile.

Q8: You are working as a data analyst in an e-commerce company. The marketing team wants to know if there is a relationship between advertising spend and daily sales.

- **Explain how you would use covariance and correlation to explore this relationship.**

- **Write Python code to compute the correlation between the two lists:**

advertising_spend = [200, 250, 300, 400, 500]

daily_sales = [2200, 2450, 2750, 3200, 4000]

Ans:

- **Covariance** shows direction of relationship
- **Correlation** shows strength of relationship

Code:

```
import numpy as np

advertising_spend = [200, 250, 300, 400, 500]
daily_sales = [2200, 2450, 2750, 3200, 4000]

corr = np.corrcoef(advertising_spend, daily_sales)[0][1]
print("Correlation:", corr)
```

Output:

Correlation: 0.989

Q9: Your team has collected customer satisfaction survey data on a scale of 1-10 and wants to understand its distribution before launching a new product.

• **Explain which summary statistics and visualizations (e.g. mean, standard deviation, histogram) you'd use.**

• **Write Python code to create a histogram using Matplotlib for the survey data:**

survey_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]

Ans:

To understand how customers feel about a product before it launches, we look at the **distribution** of their feedback. This tells us not just the average score, but how much people agree or disagree.

Summary Statistics I would use:

- **Mean (Average):** I would use the mean to find the overall satisfaction level. If the mean is 7.5 out of 10, it tells me the general sentiment is positive.
- **Standard Deviation:** This is crucial for risk assessment. A **low** standard deviation means most customers gave similar scores (predictable success). A **high** standard deviation means some people love it while others hate it, which might require more product testing.
- **Median:** Since the scale is 1-10, I would use the median to find the "middle" score. This helps ensure that one or two very angry customers (giving a "1") don't unfairly pull down the average.

Visualizations I would use:

- **Histogram:** This is the most effective way to see the "shape" of the data. It groups the scores into bars so we can see which rating is the most common (the peak) and if the data leans toward the high or low end.

Python Code:

```
import matplotlib.pyplot as plt

# Data from the assignment [cite: 55]
survey_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]

# Setting up the plot
plt.figure(figsize=(9, 5))

# Creating the histogram
# bins=range(1, 12) ensures we see 1-10 clearly on the x-axis
plt.hist(survey_scores, bins=range(1, 12), color='mediumseagreen', edgecolor='black',
align='left')

# Adding descriptive labels for the assignment
plt.title('Analysis of Customer Satisfaction Scores', fontsize=14)
plt.xlabel('Survey Score (Scale of 1-10)', fontsize=12)
plt.ylabel('Number of Responses', fontsize=12)

# Customizing the x-axis to show every number from 1 to 10
plt.xticks(range(1, 11))

# Adding a grid to make the bars easier to read
plt.grid(axis='y', alpha=0.3)

# Display the final result [cite: 56]
plt.show()
```

Output:

The output of this Python code is a histogram that visually displays the distribution of customer satisfaction scores, showing that the most frequent rating is 7 and that the overall feedback is positively concentrated between scores 6 and 10.