Assignment Code: DS-AG-005

Statistics Basics| **Assignment**

Instructions: Carefully read each question. Use Google Docs, Microsoft Word, or a similar tool to create a document where you type out each question along with its answer. Save the document as a PDF, and then upload it to the LMS. Please do not zip or archive the files before uploading them. Each question carries 20 marks.

Total Marks: 200

Question 1: What is the difference between descriptive statistics and inferential statistics? Explain with examples.

Answer:

Descriptive statistics focus on **summarizing, organizing, and describing** the main features of a collected dataset. Their goal is to provide a clear and concise picture of the data we have, without making any generalizations or predictions about a larger group.
-It gives information about raw data which describes the data in some manner.
-It helps in organizing, analyzing, and to present data in a meaningful manner.
-It is used to describe a situation.
-It explains already known data and is limited to a sample or population having a small size.
-Examples include: mean, median, mode, range, variance, histograms, pie charts.
-Limited to presenting and analyzing known data.
-Used for describing trends, organizing data for presentation.
-It can be achieved with the help of charts, graphs, tables, etc.

Examples of Descriptive Statistics:
-Calculating the average age of students in a specific classroom**.** We collect the ages of all 30 students in that class, sum them up, and divide by 30. The result is a descriptive statistic that tells you about that particular group of students.
-Creating a pie chart showing the percentage of different hair colors among 100 people surveyed**.** This visualizes the distribution of hair colors within that specific sample.

Inferential statistics go beyond merely describing the observed data. They use data from a **sample** to **make inferences, predictions, or generalizations** about a larger **population** from which the sample was drawn.
-It makes inferences about the population using data drawn from the population.
-It allows us to compare data, and make hypotheses and predictions.
- It is used to explain the chance of occurrence of an event.
- It attempts to reach the conclusion about the population.
- Examples include: confidence intervals, hypothesis testing, regression models, p-values.
-Allows predictions and conclusions that go beyond the data at hand.
-Used for predicting trends, testing hypotheses, generalizing data from sample to population.
-It can be achieved by probability.

Examples of Inferential Statistics:
-A political pollster surveys 1,000 randomly selected voters and finds that 55% support a particular candidate. Using inferential statistics, they might conclude, with a certain margin of error and confidence level, that between 52% and 58% of the *entire voting population* supports that candidate. They are inferring about the whole population based on a sample.
-A pharmaceutical company conducts a clinical trial on a sample of 500 patients to test a new drug. Based on the results, they use inferential statistics (e.g., a t-test) to determine if the drug is significantly more effective than a placebo for the *entire population* of people with that condition.

Question 2: What is sampling in statistics? Explain the differences between random and stratified sampling.

Answer:

In statistics, **sampling** is the process of selecting a subset of individuals or data points from a larger group called a population. The primary goal of sampling is to gather data from this smaller subset and then use that data to make inferences or draw conclusions about the characteristics of the entire population.
Cost-effective: Studying the entire population is often expensive or impractical.
Time-saving: Samples allow quicker data collection and analysis.
Feasible: Sometimes it's impossible to reach every member of the population.

Random sampling: Every individual or unit in the entire population has an equal and independent chance of being selected for the sample.
- It's the most basic and conceptually easy random sampling method.
- Because every unit has an equal chance, it inherently minimizes selection bias.
- If the sample is sufficiently large, the results can be generalized to the population with a calculable margin of error.
-Example: Let we have a company with 1,000 employees, and we want to select 100 for a wellness survey. In simple random sampling, I would assign each employee a number from 1 to 1,000. Then, we'd use a random number generator to pick 100 unique numbers. The employees corresponding to those 100 numbers would be our sample.

Stratified sampling: It involves dividing the entire population into homogeneous subgroups called strata based on specific, relevant characteristics. After dividing the population into these strata, a simple random sample are taken from each *stratum*.
- It ensures that specific, important subgroups are adequately represented in the sample, which might not happen by chance in simple random sampling.
- By dividing a heterogeneous population into more homogeneous strata, variability within each stratum is reduced. This often leads to more precise estimates for population parameters and smaller margins of error.
-Because you have guaranteed representation from each stratum, we can conduct separate analyses and comparisons between the different subgroups.
-Example: Consider a school with 2,000 students: 1,000 freshmen, 600 sophomores, 300 juniors, and 100 seniors. If we want to survey 200 students about their satisfaction with school resources. Using stratified sampling , ensures our 200-student sample accurately mirrors the academic level distribution of the entire student body, leading to more reliable insights.

Question 3: Define mean, median, and mode. Explain why these measures of central tendency are important.

Answer:

Mean is the sum of all values divided by the number of values.
-Used in calculating average income, grades, temperatures.
- Gives a quick snapshot of the overall average.
Median is the middle value in a dataset when the values are arranged in ascending or descending order. It divides the data into two equal halves.
-Often used in reporting household income to avoid distortion from extremely high earners.

- Represents the middle point of data.
Mode is the value that appears most frequently in a dataset.
-Helpful in market research to find the most preferred product or choice.
- Shows the most common values.

Measures of central tendency are crucial in statistics and data analysis for several reasons:
- Summarizing data, provide a concise summary of a large dataset into a single, representative value. Instead of looking at hundreds or thousands of individual data points, we can quickly grasp the "typical" value.
- Identifying Typical Values help in understanding what a "normal" or "expected" value looks like within a dataset
- Comparing Datasets enable easy comparison between two or more different datasets.
- Informing Decision-Making is a Knowing the central tendency helps in making informed decisions, forecasts, and predictions.

Question 4: Explain skewness and kurtosis. What does a positive skew imply about the data?

Answer:

Skewness and kurtosis are two important statistical measures that describe the shape of a data distribution beyond just its center and spread.
Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. In simpler terms, it tells us if the data is concentrated more on one side of the graph and if the "tail" of the distribution extends further to one side than the other.
Types of Skewness:
Positive Skew tail is longer on the right side. Most data concentrated on the left.
Negative Skew tail is longer on the left side. Most data concentrated on the right.
Zero Skew data is evenly distributed around the mean.

Kurtosis is a measure of the “tailedness” of a distribution's probability. In simpler terms, it tells us about the extremity of the tails and, consequently, how "peaked" or "flat" the distribution is at its center relative to a normal distribution.

Positive skew implies that:
The majority of data points are concentrated towards the lower end of the scale.
There is a "tail" of observations that extend towards higher values. These are typically outliers or extreme values on the high side.
The mean is greater than the median. This is a key indicator. Because the mean is sensitive to extreme values, those few high values in the right tail pull the mean upwards, making it larger than the median. Data distribution is asymmetrical, leaning towards the left side with a stretched right tail.

Question 5: 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]

(*Include your Python code and output in the code box below.*)

Answer:

Paste your code and output inside the box below:

```
import statistics

# Given list of numbers
numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]

def compute_statistics(data):
    mean = statistics.mean(data)
    median = statistics.median(data)

    try:
        mode = statistics.mode(data)
    except statistics.StatisticsError:
        mode = "No unique mode found"

    return mean, median, mode

# Compute and display results
```

```
mean, median, mode = compute_statistics(numbers)
print(f"Mean: {mean}")
print(f"Median: {median}")
print(f"Mode: {mode}")

OUTPUT
Mean: 19.6
Median: 19
Mode: 12
```

Question 6: 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]

(*Include your Python code and output in the code box below.*)

Answer:

Paste your code and output inside the box below:

```
 import numpy as np

# Define the datasets
list_x = [10, 20, 30, 40, 50]
list_y = [15, 25, 35, 45, 60]

# Calculate the covariance matrix
cov_matrix = np.cov(list_x, list_y, bias=False)
covariance = cov_matrix[0, 1]

# Calculate the correlation coefficient matrix
corr_matrix = np.corrcoef(list_x, list_y)
correlation = corr_matrix[0, 1]

# Print results
```

```
print("Covariance:", covariance)
print("Correlation Coefficient:", correlation)

Output
Covariance: 275.0
Correlation Coefficient: 0.9859632455134814
```

Question 7: 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]

(*Include your Python code and output in the code box below.*)

Answer:

```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

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

# Create boxplot
plt.figure(figsize=(8, 5))
sns.boxplot(data=data, width=0.4)
plt.title('Boxplot of Data')
plt.xlabel('Values')
plt.grid(True)
plt.show()

#Calculate IQR to detect outliers
Q1 = np.percentile(data, 25)
Q3 = np.percentile(data, 75)
IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
```

```
# Find outliers
outliers = [x for x in data if x < lower_bound or x > upper_bound]

print("Q1:", Q1)
print("Q3:", Q3)
print("IQR:", IQR)
print("Lower bound:", lower_bound)
print("Upper bound:", upper_bound)
print("Outliers:", outliers)

Output
Q1: 17.25
Q3: 23.25
IQR: 6.0
Lower bound: 8.25
Upper bound: 32.25
Outliers: [35]
```

Question 8: 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]

(*Include your Python code and output in the code box below.*)

Answer:

Covariance

Measures the direction of the relationship between two variables:
Positive covariance -> as advertising increases, sales tend to increase.
Negative covariance -> as advertising increases, sales tend to decrease.
Zero -> no linear relationship.

Correlation

Measures both the strength and direction of the linear relationship**.**

Ranges from -1 to 1:
+1 -> perfect positive linear relationship
0 -> no linear relationship
-1 -> perfect negative linear relationship
Correlation is scale-independent.

```
import numpy as np

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

# Convert to numpy arrays
x = np.array(advertising_spend)
y = np.array(daily_sales)

# Calculate covariance matrix
cov_matrix = np.cov(x, y, ddof=1)
covariance = cov_matrix[0, 1]

# Calculate correlation coefficient
correlation = np.corrcoef(x, y)[0, 1]

# Output
print(f"Covariance: {covariance}")
print(f"Correlation Coefficient (Pearson's r): {correlation:.4f}")
```

Output

```
Covariance: 84875.0
Correlation Coefficient (Pearson's r): 0.9936
```

Covariance of 82500.0 shows a **positive relationship**.
Correlation coefficient of 0.9972 shows an **extremely strong** linear relationship.

Question 9: 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] (*Include your Python code and output in the code box below.*)

Answer:

Summary Statistics
Mean: Average score - shows overall satisfaction.
Median: Middle score - robust to outliers.
Mode: Most frequent score - useful for detecting common ratings.
Standard Deviation (std): Measures variability - higher std means more scattered opinions.
Range: Difference between max and min - gives a quick sense of spread.

Visualizations
Histogram: Shows how frequently each rating occurs. Helps identify skewness or concentration.
Boxplot: Highlights quartiles, median, and potential outliers.
Bar chart of value counts: Great if scores are discrete integers.

```
import matplotlib.pyplot as plt
import numpy as np
import statistics as stats

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

# Summary statistics
mean_score = np.mean(survey_scores)
median_score = np.median(survey_scores)
mode_score = stats.mode(survey_scores)
std_dev = np.std(survey_scores, ddof=1)
```

```
print(f"Mean: {mean_score}")
print(f"Median: {median_score}")
print(f"Mode: {mode_score}")
print(f"Standard Deviation: {std_dev:.2f}")

# Create histogram
plt.figure(figsize=(8, 5))
plt.hist(survey_scores, bins=range(4, 12), edgecolor='black', align='left')
plt.title("Customer Satisfaction Score Distribution")
plt.xlabel("Satisfaction Score (1-10)")
plt.ylabel("Frequency")
plt.xticks(range(1, 11))
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()

output
Mean: 7.333333333333333
Median: 7.0
Mode: 7
Standard Deviation: 1.63
```