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 summarize and describe the main features of a dataset, while inferential statistics use sample data to make predictions and generalizations about a larger population. Descriptive statistics involve measures like mean, median, and standard deviation, describing the "what" of the data, whereas inferential statistics involve techniques like hypothesis testing and confidence intervals to understand the "why" and "what next" from data about a larger, unstudied group.

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

Answer: Sampling is selecting a subset of individuals or items from a larger population to represent the whole group for analysis. Simple random sampling involves selecting participants at random from the entire population, ensuring every member has an equal chance of selection. Stratified sampling first divides the population into distinct, non-overlapping groups called strata (e.g., by age or gender), and then randomly selects participants from each stratum. The key difference is that random sampling selects from the whole population, while stratified sampling first categorizes the population into subgroups before randomly sampling from those subgroups, which improves precision and provides specific insights into subgroups.

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

Answer:

**Mean**: The arithmetic mean is found by summing all the numbers in a dataset and dividing by the total count of numbers.

**Median**: The median is the middle value in a dataset that has been ordered from smallest to largest. If there is an even number of data points, the median is the average of the two middle values.

**Mode**: The mode is the value that appears most often in a dataset. A dataset can have more than one mode or no mode at all.

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

Answer: Skewness measures a distribution's lack of symmetry, indicating if it leans to the left or right, while kurtosis measures the "tailedness" or peakedness of the distribution's tails compared to a normal distribution. A positive skew implies that the data is concentrated on the lower values, with a longer, heavier right tail and the mean being greater than the median.

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

import numpy as np

np.mean(numbers)

np.float64(19.6)

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

import numpy as np

np.median(numbers)

np.float64(19.0)

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

import statistics as st

st.mode(numbers)

12

Paste your code and output inside the box below:

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: import math

list\_x = [10, 20, 30, 40, 50]

list\_y = [15, 25, 35, 45, 60]

n = len(list\_x)

# Mean

mean\_x = sum(list\_x) / n

mean\_y = sum(list\_y) / n

# Covariance (sample, using denominator n or using n‑1 depending on definition — I'll use population version with n)

cov\_xy = sum((x - mean\_x)\*(y - mean\_y) for x,y in zip(list\_x, list\_y)) / n

# Standard deviations (population)

std\_x = math.sqrt(sum((x - mean\_x)\*\*2 for x in list\_x) / n)

std\_y = math.sqrt(sum((y - mean\_y)\*\*2 for y in list\_y) / n)

correlation = cov\_xy / (std\_x \* std\_y)

mean\_x, mean\_y, cov\_xy, std\_x, std\_y, correlation

(30.0, 36.0, 220.0, 14.142135623730951, 15.620499351813308, 0.9958932064677039)

Paste your code and output inside the box below:

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 numpy as np

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

plt.figure(figsize=(8, 6))

plt.boxplot(data)

plt.title('Boxplot of Numeric Data')

plt.ylabel('Value')

plt.grid(True)

plt.show()

# Calculate quartiles and IQR to identify 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

outliers = [x for x in data if x < lower\_bound or x > upper\_bound]

print(f"First Quartile (Q1): {Q1}")

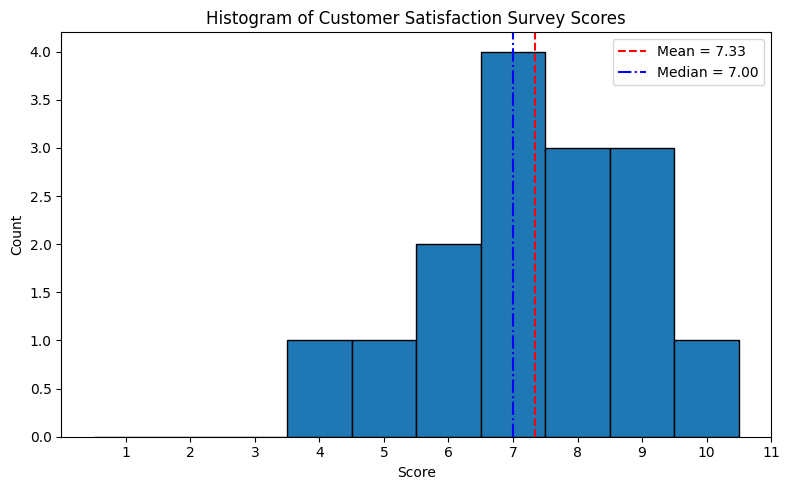
print(f"Third Quartile (Q3): {Q3}")

print(f"Interquartile Range (IQR): {IQR}")

print(f"Lower Bound for Outliers: {lower\_bound}")

print(f"Upper Bound for Outliers: {upper\_bound}")

print(f"Identified Outliers: {outliers}")

11

First Quartile (Q1): 17.25

Third Quartile (Q3): 23.25

Interquartile Range (IQR): 6.0

Lower Bound for Outliers: 8.25

Upper Bound for Outliers: 32.25

Identified 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: import numpy as np

from scipy import stats

advertising\_spend = [200, 250, 300, 400, 500]

daily\_sales = [2200, 2450, 2750, 3200, 4000]

#use numpy

x = np.array(advertising\_spend)

y = np.array(daily\_sales)

# compute covariance (sample covariance)

cov\_matrix = np.cov(x, y, ddof=1) # ddof=1 → sample covariance

cov\_xy = cov\_matrix[0, 1]

print("Covariance between ad spend and sales:", cov\_xy)

# compute correlation via numpy

corr\_matrix = np.corrcoef(x, y)

corr\_xy = corr\_matrix[0, 1]

print("Pearson correlation coefficient (numpy):", corr\_xy)

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: import matplotlib.pyplot as plt

import numpy as np

import statistics

# Your data

survey\_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]

# Summary statistics

n = len(survey\_scores)

mean\_ = statistics.mean(survey\_scores)

median\_ = statistics.median(survey\_scores)

stdev\_ = statistics.stdev(survey\_scores)

mn = min(survey\_scores)

mx = max(survey\_scores)

q1, q3 = np.percentile(survey\_scores, [25, 75])

print(f"n = {n}")

print(f"Mean = {mean\_:.2f}, Median = {median\_:.2f}")

print(f"Std dev = {stdev\_:.2f}")

print(f"Min = {mn}, Q1 = {q1:.2f}, Q3 = {q3:.2f}, Max = {mx}")

# Plot histogram

plt.figure(figsize=(8, 5))

# Choose number of bins — here we use 7 bins (or you can try other heuristics)

bins = range(1, 12)  # from 1 to 11 so that 10 is included as a full bin

plt.hist(survey\_scores, bins=bins, edgecolor='black', align='left')

# Overlay lines for mean and median

plt.axvline(mean\_, color='red', linestyle='--', linewidth=1.5, label=f"Mean = {mean\_:.2f}")

plt.axvline(median\_, color='blue', linestyle='-.', linewidth=1.5, label=f"Median = {median\_:.2f}")

plt.title("Histogram of Customer Satisfaction Survey Scores")

plt.xlabel("Score")

plt.ylabel("Count")

plt.xticks(bins)  # show integer ticks

plt.legend()

plt.tight\_layout()

plt.show()

n = 15

Mean = 7.33, Median = 7.00

Std dev = 1.63

Min = 4, Q1 = 6.50, Q3 = 8.50, Max = 10