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 vs Inferential Statistics
Descriptive Statistics and Inferential Statistics are 2 main branches of statistics both have different purpose.
Descriptive Statistics: Descriptive statistics include methods for summarizing and describing the important features of a dataset it provides simple summaries about the sample and the measures.
Purpose is to organize and summarize data in a meaningful way to understand it more easily.
Techniques: Measures of central tendency (mean, median, mode)….measures of variability (range, variance, standard deviation)……Charts and graphs (histograms, pie charts, bar charts)
eg: test scores of 100 students..
Mean: Average score
Median: Middle score when sorted
Standard deviation: How spread out the scores are
Suppose scores are: 65, 70, 75, 80, 85, the mean would be 75 and the standard deviation telling about you how varied the scores are.

Inferential Statistics:: Inferential statistics use sample of data to make generalizations or inferences about a larger population it helps to make predictions, test hypotheses, or estimate population parameters….used to make conclusions about a population based on a sample.
Techniques: Hypothesis testing (t-tests, chi-square tests), Confidence intervals, Regression analysis
eg:
randomly selecting 100 students from a school of 1000 studnts based on test scores we want to infer the average score for the entire school
taking a random sample of 100 students and calculating the average score eg 72
Using inferential statistics we can estimate that the average test score for all 1000 students is around 72, within a specific margin of error.

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

Answer:

Sampling is the process of selecting a group called a sample from a larger population to make conclusions or predictions about the entire population instead of studying everyone in a population we just study a small part of it and use the results to make guesses about the whole.
eg: if we want to know the average height of all students in a school, we will not measure all students instead we will select a few students (sample) and calculate the average height based on them.

random sampling vs stratified sampling
Random Sampling: in random sampling, every person in the population has equal chance of being selected it works like randomly picking people/items and no specific group or characteristic is considered when choosing.
eg: selecting 10 students from a school of 1000 we will put all names in a hat and draw 10 names randomly each student has equal chance of being selected. It's easy to do and there is no bias in choosing the sample but sometimes it can miss important groups if the population is very diverse.

Stratified Sampling: the population is divided into subgroups called strata that share a common characteristic and then we randomly sample from each subgroups it makes sure that every subgroup is represented it works like divide the population into different groups (strata) based on specific traits then randomly select people from each group.
eg:
In a school with 60 boys and 40 girls we want to select 10 students instead of just randomly choosing 10 students we divide students into 2 groups ie. boys and girls then we randomly pick a certain number of boys and girls so both groups are represented this way it is more accurate because it ensures every group is represented and it is good for diverse populations but it is a little more complicated and takes more time than random sampling and we need to know the characteristics of the population before sampling.
In summary\: random Sampling is simple and gives each person an equal chance of being selected. stratified sampling is more precise because it ensures every subgroup is represented in the sample and sampling helps in saving time and resources but the method you choose depends on the nature of the population.

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

Answer:

Mean is the average of a set of numbers we can find it by adding all the numbers together and then dividing by how many numbers there are.
eg:
2, 4, 6, 8 : Mean=2+4+6+8/4 =5
mean is 5

Median: median is the middle number in a set of numbers when they are arranged in order from smallest to largest if there is an even number of values the median is the average of the 2 middle numbers.
eg:
2, 4, 6, 8
arrange numbers in order: 2, 4, 6, 8.
There are 4 numbers, so take the middle 2: 4 … 6
median is the average of these two ie 4+6/2=5
median is 5
suuppose 1, 3, 5 then median is 3, because it's the middle number

Mode: mode is the number that appears most often in a set of numbers.
eg:
1, 2, 2, 3, 3, 3, 4….. number 3 appears three times more than any other number so 3 is the mode.
Situation like 1, 1, 2, 2, 3, 3 this is a bimodal distribution it has two modes: 1 and 2.
If no number repeats then there is no mode.

Why are these measures of central tendency important:

mean, median, mode give a quick summary of a dataset they help to understand the central point around which the data is spread.
Instead of looking at every single data point these measures allow us to represent the data with a single number making it easier to understand and compare datasets.
Mean is useful when the data is evenly distributed ie. no extreme values
Median is useful when there are extreme values or outliers as it’s not affected by very high or low values.
Mode is useful when we need to find the most common value……these measures are used to make important decisions based on data.
these measures help to spot trends and patterns in the data eg: if the mean and median are very different it could mean that there are some outliers affecting the data.

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

Answer:

Skewness measure the asymmetry of a data distribution around its mean.
symmetric distribution has a skewness of 0 like a normal distribution.
positive skew (right skewed) means the tail on right side is longer or fatter than the left in this condition mean is greater than median.
negative skew (left skewed) means tail on left side is longer or fatter than the right meaning that mean is less than the median.

A positive skew implies that most data values are concentrated on the lower end but a few unusually high values pull the mean to the right eg. this occurs in income data where most people earn average salaries but a few high earners raise the average.

Kurtosis measures the tailedness or peakness of a distribution like how heavily the tails differ from the normal distribution. 3 types of them are below:
Mesokurtic (kurtosis appro ~ 3) : Normal distribution
Leptokurtic (kurtosis > 3): More peaked with heavier tails..higher chance of extreme values ie outliers
Platykurtic (kurtosis < 3): Flatter peak with lighter tails few outliers

High kurtosis indicates a greater likelihood of extreme outcomes
Low kurtosis suggests a more uniform and flatter distribution

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

numbers = [12, 15, 12, 18, 19, 12, 20, 22, 19, 19, 24, 24, 24, 26, 28]
mean = statistics.mean(numbers)
median = statistics.median(numbers)
mode = statistics.mode(numbers)

print("mean is", mean)
print("median is", median)
print("mode is", mode)
```

o/p

```
mean is 19.6
median is 19
mode is 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

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

x = np.array(list_x)
y = np.array(list_y)

cov_matrix = np.cov(x, y, bias=False)
covariance = cov_matrix[0][1]

correlation_matrix = np.corrcoef(x, y)
```

```
correlation = correlation_matrix[0][1]

print("covariance is", covariance)
print("correlation coefficient is", correlation)


o/p

covariance is 275.0
correlation coefficient is 0.995893206467704
```

Question 7: Write a Python script to draw a boxplot for the following numeric list and identify its outliers. Explain the result:

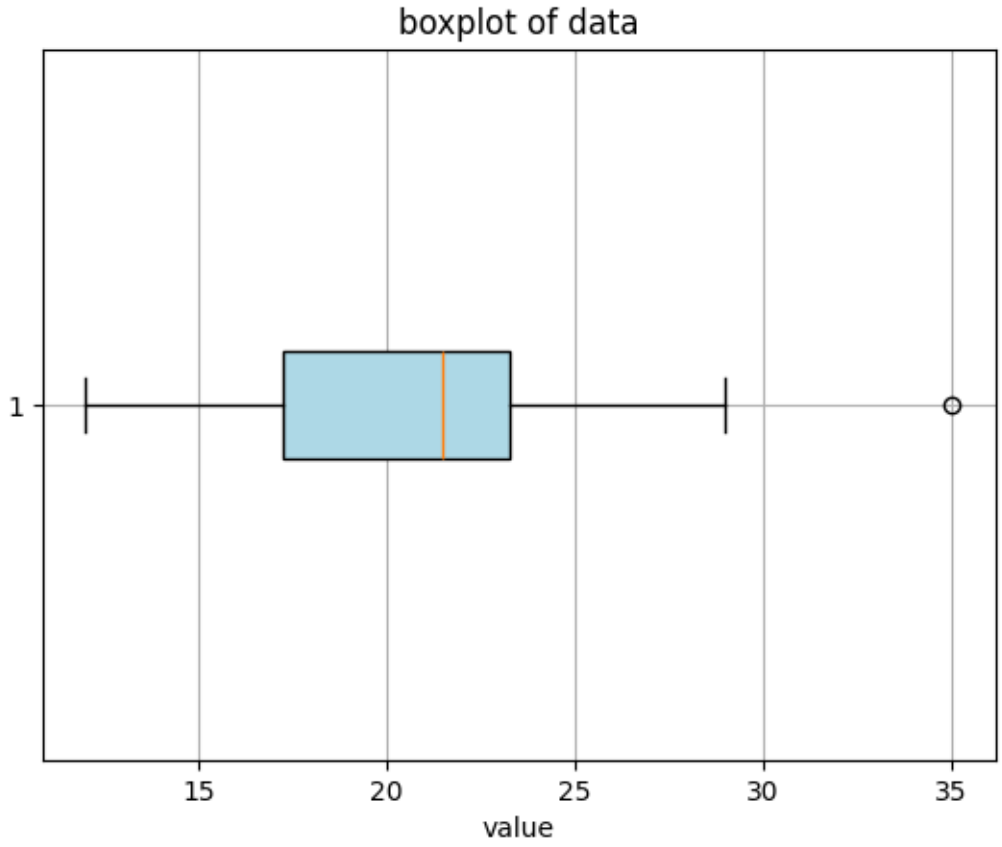
data = [12, 14, 14, 15, 18, 19, 19, 21, 22, 22, 23, 23, 24, 26, 29, 35]

(*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.boxplot(data, vert=False, patch_artist=True, boxprops=dict(facecolor='lightblue'))
plt.title("boxplot of given data")
plt.xlabel("value")
plt.grid(True)
plt.show()

# identifying outliers manually using IQR method
Q1 = np.percentile(data, 25)  # first quartile
Q3 = np.percentile(data, 75)  # third quartile
IQR = Q3 - Q1

# outlier thresholds
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# finding 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)
```



Q1: 17.25
Q3: 23.25
IQR: 6.0
Lower Bound: 8.25
Upper Bound: 32.25
Outliers: [35]

Explanation pls check here:
IQR is 6.0.
Using IQR rule:
Lower threshold = Q1 − 1.5×IQR = 9.0
Upper threshold = Q3 + 1.5×IQR = 33.0

only value above 33.0 is 35 so it is classified as outlier

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:

To find the relationship between advertising spend and daily sales we can use:
Covariance: it shows whether the two variables move in the same or opposite direction:
Positive covariance: as advertising increases sales tend to increase
Negative covariance: as advertising increases sales tend to decrease but covariance cannot tell us the strength of the relationship or standardize the values.

Correlation coefficient: it measures the strength and direction of the linear relationship between two variables.
Ranges from -1 to +1
+1: perfect positive relationship
0: no linear relationship
−1: perfect negative relationship
high positive correlation tells that higher advertising spend is strongly associated with higher daily sales.

Code

```
import numpy as np

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

```
x = np.array(advertising_spend)
y = np.array(daily_sales)
cov_matrix = np.cov(x, y, bias=False)
covariance = cov_matrix[0][1]

correlation_matrix = np.corrcoef(x, y)
correlation = correlation_matrix[0][1]

print("covariance is", covariance)
print("correlation coefficient is", correlation)


o/p

covariance is 84875.0
correlation coefficient is 0.9935824101653329
```

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:

To understand the distribution of customer satisfaction survey scores scale 1–10 we should use:
Summary Statistics:

- Mean: it shows the average satisfaction score
- Median: it shows the central tendency ie. less sensitive to outliers
- Mode: it identifies most common score
- Standard deviation (sd): it measures how spread out the scores are around mean
- Skewness & Kurtosis: helps t0 understand asymmetry and peakedness of the distribution

Visualizations:

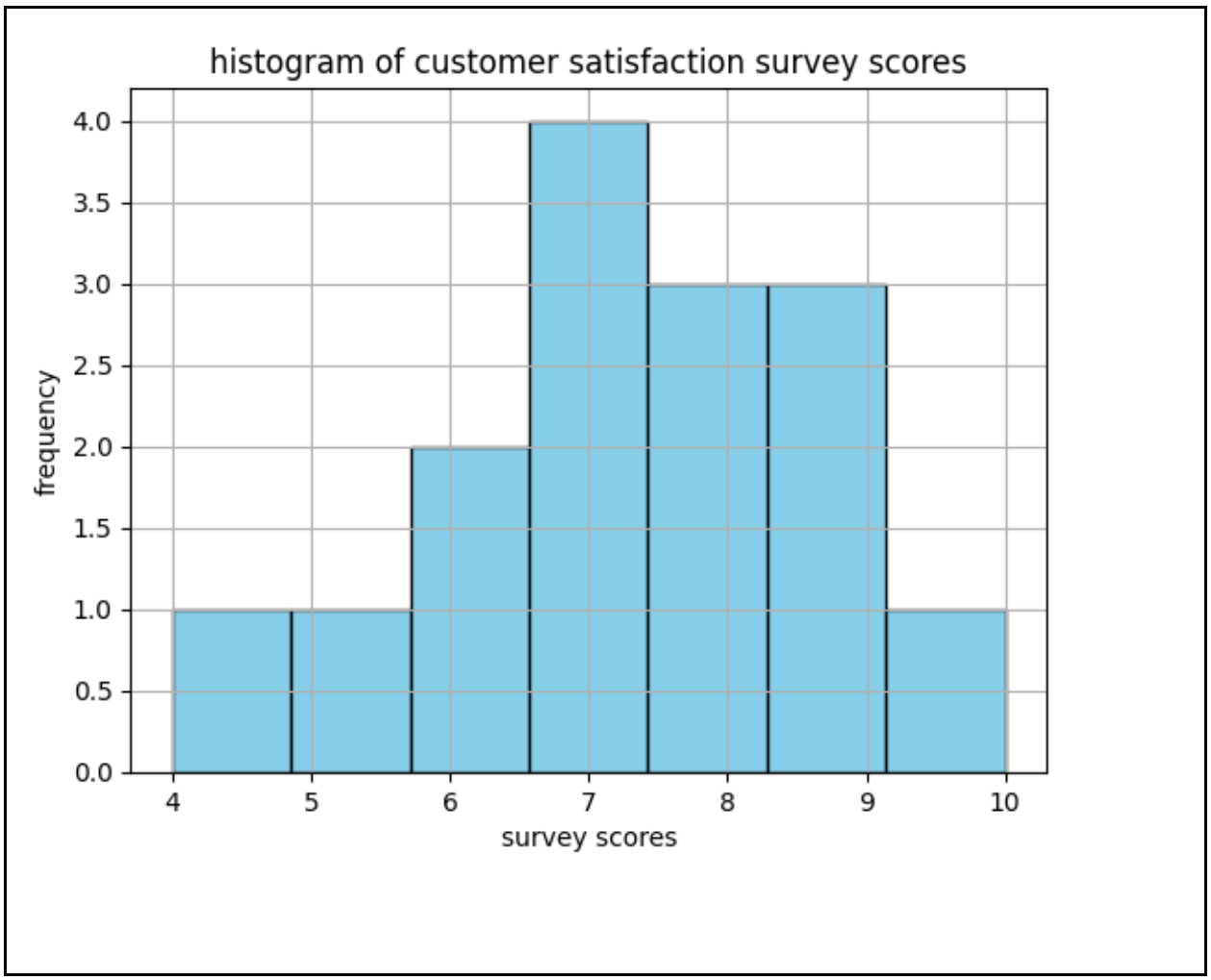
- Histogram: it displays frequency distribution of scores helping to see if the data is symmetric, skewed or bimodal

Code

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

survey_scores = [7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]
mean = statistics.mean(survey_scores)
median = statistics.median(survey_scores)
mode = statistics.mode(survey_scores)
stand_dev = statistics.stdev(survey_scores)

print("mean:", mean)
print("median:", median)
print("mode:", mode)
print("standard devation:", stand_dev)

plt.hist(survey_scores, bins=7, edgecolor='black', color='skyblue')
plt.title("histogram of customer satisfaction survey scores")
plt.xlabel("survey scores")
plt.ylabel("frequency")
plt.grid(True)
plt.show()
```

o/p

```
mean: 7.333333333333333
median: 7
mode: 7
standard devation: 1.632993161855452
```


histogram of customer satisfaction survey scores
frequency
survey scores
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4
5
6
7
8
9
10