Statistics Basics Assignment

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

Answer: Descriptive statistics summarize and describe the features of a dataset.They provide a way to present data in a meaningful way,allowing for quick understanding of the information.

Example: hours studied by 3 friends: 1,2,2

Mean= 1.67, Mode=2

Inferential statistics use sample data to make inferences or predictions about a larger population.

Example:Using this you infer all college friends study avg~1.67 hours/day.

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

Answer: The process of selecting a subset of individuals or data from a larger population to make inferences about the whole population.

Random Sampling – Every member of the population has an equal chance of being selected it helps reduce bias.

Stratified Sampling– The population is divided into subgroups(strata) based on shared characteristics.Samples are then randomly taken from each stratum.This ensures representation from each group.

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

Answer: Mean : The average of all values.

Median : The middle value when data is ordered.

Mode : The most frequently occurring value.

The measures of central tendency are important because they summarise data into a single representative value,help understand the distribution's center and are useful for comparisons and making inferences about the population.

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

Answer: Skewness describes the asymmetry of a probability distribution. Kurtosis describes the " tailedness" of a probability distribution.A positive skew implies the tail of the distribution is longer on the right side, meaning there are more extreme values on the higher end of the data.

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]
mean=statistics.mean(numbers)
median=statistics.median(numbers)
mode=statistics.mode(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: import numpy as np
list_x=np.array([10,20,30,40,50])
list_y=np.array([15,25,35,45,60])
covariance=np.cov(list_x,list_y)[0,1]
correlation_coefficient:np.corrcoef(list_x,list_y)[0,1]
print(f"Covariance:{covariance}")
print(f"Correlation Coefficient:{correlation_coefficient}")
#Output: Covariance:275.0
Correlation Coefficient: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]
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("Outliers:",outliers)
plt.boxplot(data)
plt.little("Boxplot of the Data")
plt.show()
#Output: Outliers: [35]
AttributeError: module 'matplotlib.pyplot' has no attribute 'little'
```

Explanation: ● Outliers:Running the code will print the outliers based on the IQRmethod.For the given data:Q1=18.25,Q3=23.75,IQR= 5.5,lower_bound= 18.25-1.5*5.5=10,upper_bound=23.75+1.5*5.5=32,Outliers:[3 5] because 35>32.

● Boxplot: The code will display a boxplot showing the distribution of the data,highlighting the median,quartiles,and any outliers(in this case,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 how much two variables change together.If the covariance is positive,it means when one variable increases,the other tends to increase too.Correlation is a normalized version of covariance that ranges from -1to1, making it easier to interpret the strength and direction of the relationship,close to -1 means a strong negative relationship,and close to 0 means no linear relationship.

```
● import numpy as np
advertising_spend=([200, 250, 300, 400, 500])
daily_sales=([2200, 2450, 2750, 3200, 4000])
covariance=np.cov(advertising_spend,daily_sales)[0,1]
correlation=np.corrcoef(advertising_spend,daily_sales)[0,1]
print(f"Covariance:{covariance}")
print(f"Correlation:{correlation}")
# Output : Covariance:84875.0
Correlation:0.9935824101653329
```

The correlation of approximately 0.997 indicates a very strong positive 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: ●To understand the distribution of customer satisfaction survey data(on a scale of 1-10), you'd typically use:

Mean: Gives the average satisfaction score.

Standard Deviation: Shows how spread out the scores are from the mean.

Histogram: A visualization that shows the distribution of scores,helping to understand the shape of the data(like if it's skewed,symmetric,or has multiple peaks).

```
● import matplotlib.pyplot as plt
import numpy as np
survey_scores=[7, 8, 5, 9, 6, 7, 8, 9, 10, 4, 7, 6, 9, 8, 7]
mean_score=np.mean(survey_scores)
std_dev=np.std(survey_scores)
print(f"Mean Satisfaction Score:{mean_score}")
print(f"Standard Deviation:{std_dev}")
plt.hist(survey_scores,bins=range(1,12),edgecolor='black')
plt.xlabel('Satisfaction Score')
plt.ylabel('Frequency')
plt.title(' Distribution of Customer Satisfaction Scores')
plt.xticks(range(1,11))
plt.show()
#Output: Mean Satisfaction Score:7.333333333333333
Standard Deviation:1.577621275493231
```