Foundations of statistics for Data analytics and machine learning with Excel

Assignment-1: Basic Statistics

Name: _____ **Id:** _____

Instructions

- Justify you answers for quiz questions also.
- Explain answers in detail for any question.

Introduction to Statistics

Part A: Quiz

1. What is the primary purpose of descriptive statistics?
   A. To predict future outcomes
   B. To draw conclusions from sample data
   C. To summarize and describe data
   D. To perform hypothesis testing
2. Which of the following is a measure of central tendency?
   A. Range
   B. Mean
   C. Variance
   D. Standard deviation
3. In a perfectly symmetrical distribution, which of the following is true?
   A. Mean > Median > Mode
   B. Mean < Median < Mode
   C. Mean = Median = Mode
   D. Median = Mode $\neq$ Mean
4. The interquartile range (IQR) is defined as:
   A. The difference between the maximum and minimum values
   B. The average of the first and third quartiles
   C. The difference between the first and third quartiles
   D. The sum of all quartiles
5. Which of the following best describes a population in statistics?
   A. A numerical summary of a sample

B. A subset of a sample

C. All possible observations of interest

D. A histogram of the sample data

6. Which type of graph is best suited to compare the frequencies of different categories?

   A. Histogram

   B. Box plot

   C. Bar plot

   D. Scatter plot

7. A histogram is primarily used to:

   A. Show the relationship between two variables

   B. Display the distribution of a numerical variable

   C. Compare proportions between categories

   D. Summarize central tendency and spread

8. In a box plot, the length of the box represents:

   A. The mean of the data

   B. The total range

   C. The interquartile range (IQR)

   D. The median

9. Which of the following graphs is most appropriate to visualize parts of a whole?

   A. Histogram

   B. Pie chart

   C. Scatter plot

   D. Box plot

10. Scatter plots are most useful for:

    A. Showing frequency distributions

    B. Analyzing categorical data

    C. Displaying trends in time series

    D. Investigating relationships between two quantitative variables

11. If you want to detect potential outliers in a dataset, which graph would be most appropriate?

    A. Bar plot

    B. Pie chart

    C. Box plot

    D. Histogram

12. When would it be inappropriate to use a pie chart?

    A. When comparing parts of a whole

B. When dealing with a small number of categories

C. When the total percentage exceeds 100%

D. When working with nominal data

13. A histogram differs from a bar plot in that:

A. It uses non-adjacent bars

B. It displays categorical data

C. It shows the frequency of numeric intervals

D. It summarizes proportions of a whole

14. Which of the following would best help you assess correlation?

A. Histogram

B. Bar plot

C. Scatter plot

D. Box plot

15. What does the height of a bar represent in a bar chart?

A. The variance of the category

B. The proportion or frequency of that category

C. The median value of the category

D. The correlation with other categories

16. Which of the following is a characteristic of **descriptive statistics** ?

A. Makes predictions about a population

B. Uses charts and graphs to summarize data

C. Draws conclusions from sample data

D. Computes p-values and confidence intervals

17. **Inferential statistics** involves:

A. Organizing and displaying data

B. Calculating the median of a dataset

C. Using sample data to make generalizations about a population

D. Computing the range and interquartile range

18. Which of the following best illustrates inferential statistics?

A. A histogram of student test scores

B. The mean salary of 100 employees

C. Predicting election results using a poll

D. A bar chart of favorite ice cream flavors

19. What is a **population** in statistics?

A. A small group selected for analysis

B. A numerical value calculated from a sample

C. The entire group of individuals or items under study

D. A list of selected data values

20. A **sample** is:

A. Always larger than a population

B. A summary statistic

C. A subset taken from a population

D. The total number of measurements in a study

21. Which of the following best describes the relationship between a sample and a population?

A. The sample replaces the population

B. The sample is always the same as the population

C. The sample provides estimates about the population

D. The population is selected based on the sample

22. Which of the following is a reason to use a sample instead of a population?

A. It increases the complexity of the study

B. It eliminates the need for statistics

C. Studying the entire population is often impractical

D. It always gives the exact value for the population

23. Which of the following is an example of **nominal data** ?

A. Temperature

B. Student ID numbers

C. Eye color

D. Ranking in a race

24. **Ordinal data** is different from nominal data because:

A. It has a meaningful zero

B. It represents measured quantities

C. It can be meaningfully ordered

D. It cannot be categorized

25. Which of the following is an example of **discrete numerical data** ?

A. Height in centimeters

B. Number of cars in a parking lot

C. Weight of a watermelon

D. Temperature in Fahrenheit

26. **Continuous data** differs from discrete data in that:

A. It can take any value within a range

B. It only includes whole numbers

C. It cannot be measured

D. It is only used in business

27. Which of the following best describes **categorical data** ?

A. Data that can be measured with units

B. Data expressed in ranges

C. Data that describes qualities or characteristics

D. Data used only in experiments

28. A school conducts a survey asking students to select their favorite subject: Math, English, Science, or History. What type of data is being collected?

A. Ordinal

B. Nominal

C. Discrete

D. Continuous

29. A teacher records the number of books each student read over summer break. What type of data is this?

A. Nominal

B. Ordinal

C. Discrete

D. Continuous

30. Students rate the cafeteria food as: Poor, Fair, Good, Excellent. This is an example of:

A. Continuous data

B. Nominal data

C. Ordinal data

D. Discrete data

31. A fitness app records the time (in minutes) each person exercises daily. What type of data is this?

A. Discrete

B. Nominal

C. Ordinal

D. Continuous

32. A survey asks for students' birth months. What kind of data is this?

A. Nominal

B. Ordinal

C. Discrete

D. Continuous

Part B: Short Answer

1. In your own words, define **descriptive statistics** .
2. In your own words, define **inferential statistics** .
3. Give one example of descriptive statistics and one example of inferential statistics.

Part C: Fill in the Table

Instructions: Complete the table comparing descriptive and inferential statistics.

Aspect	**Descriptive Statistics**	**Inferential Statistics**
Purpose		
Uses		
Examples		
Data Scope		

Population and Sample

Quiz

1. Which of the following best defines a **population** in statistics?
    A. A small group of individuals chosen from a dataset
    B. A collection of numerical variables
    C. The entire group you want to draw conclusions about
    D. A fixed set of outcomes
2. A **sample** is:
    A. The entire group being studied
    B. The average of the population
    C. A subset of the population used to make inferences
    D. Always larger than the population
3. Which of the following is an example of a **population** ?
    A. All students in your school
    B. 100 students selected from your school
    C. One classroom of students
    D. Your group of friends
4. Why do we often use **samples** in statistics?
    A. Because samples give biased results
    B. Because populations are usually too small
    C. Because it's often impractical to collect data from the entire population
    D. Because we always know the population mean

Short answer questions

1. Define **population** in your own words.
2. Define **sample** in your own words.
3. Give one real-life example of a population and a corresponding sample.

Fill in the Comparison Table

Instructions: Fill in the table below comparing populations and samples.

Aspect	Population	Sample
Definition		
Size		
Example		
Statistic/Parameter		

Reasoning and Application

1. A teacher wants to understand the average number of hours students in a district study each week. Should she study the entire population or a sample? Justify your answer.

Data Types

Short Answer questions

1. Define **nominal** and **ordinal** data. Give one example of each.
2. Define **discrete** and **continuous** numerical data. Give one example of each.
3. Explain one key difference between categorical and numerical data.
4. A teacher ranks students based on performance as 1st, 2nd, and 3rd. What type of data is this? Why?
5. In a class survey, students are asked how many siblings they have. What type of data is this? Explain.
6. Students were asked to write down their heights in centimeters. What type of data is this? Why?

Complete the Table

Instructions: Complete the table with definitions and examples of different data types.

Data Type	Definition	Example
Nominal		
Ordinal		
Discrete		
Continuous		

Categorize the Data

Instructions: For each scenario below, identify the data type (nominal, ordinal, discrete, or continuous) and explain your reasoning.

1. Recording students' blood types (A, B, AB, O)
   Type: _____
   Reason: _____

2. Counting the number of pencils in each student's pencil case
   Type: _____
   Reason: _____

3. Students rate how stressed they feel on a scale from 1 (low) to 5 (high)
   Type: _____
   Reason: _____

4. Measuring the daily temperature in degrees Celsius
   Type: _____
   Reason: _____

Mean, Median and Mode

Quiz

1. The **mean** is:
   A. The middle value when the data is arranged
   B. The most frequent value
   C. The sum of values divided by the number of values
   D. The highest value in the dataset

2. The **median** is:
   A. The average of all numbers
   B. The most repeated number
   C. The number in the middle when arranged
   D. The range of the data

3. When a data set has two modes, it is called:
   A. Unimodal
   B. Bimodal
   C. Multimodal
   D. No mode

4. Which measure of central tendency is **not affected** by extreme values?
   A. Mean
   B. Median

C. Mode

D. Range

5. The mode of the dataset: 5, 6, 6, 7, 7, 7, 8, 9 is:

   A. 6

   B. 7

   C. 8

   D. 9

Short Answer – Conceptual Understanding

1. Explain the difference between the **mean** and the **median**.
2. When is the **median** a better choice than the **mean** to describe data?
3. Can a data set have more than one mode? Explain.
4. How is the presence of outliers identified using mean and median ?

Part C: Numerical Problems

1. Find the mean, median, and mode of the following data:
   12, 15, 13, 12, 16, 14, 13, 12
2. A class has the following test scores:
   72, 75, 78, 72, 80, 85, 90, 72, 95
   Calculate the mode and explain why the mode is useful in this case.
3. The following are ages (in years) of participants in a study:
   18, 21, 22, 19, 18, 35, 22, 18, 23
   a) Find the mean age
   b) Find the median age
   c) Find the mode

Variance and Standard deviation

Quiz

1. What does **variance** measure in a dataset?

   A. The middle value

   B. The most frequent value

   C. The spread of the data from the mean

   D. The average value

2. The **standard deviation** is:

   A. The square root of the variance

   B. The square of the mean

C. Always smaller than the mean

D. A measure of the average

3. If the standard deviation is **0** , it means:

   A. The data is very spread out

   B. All values are different

   C. There are extreme outliers

   D. All data values are the same

4. A **larger standard deviation** implies:

   A. Data values are closer to the mean

   B. The mean is higher

   C. The data is more spread out

   D. All data points are equal

Short Answer questions

1. Explain the difference between **variance** and **standard deviation**.
2. Why do we square the differences from the mean when calculating variance?
3. A dataset has a high standard deviation. What does that tell you about the data?

Numericals

1. Find the variance and standard deviation for the data set:
   2, 4, 6, 4, 4
2. A student scores the following marks in 5 subjects: 80, 85, 75, 90, 70.
   Calculate the mean and standard deviation.
3. Two classes have test score data with the same mean but different standard deviations. What does this suggest about the consistency of scores?

Skewness and Kurtosis

Quiz

1. A distribution with a **long right tail** is:

   A. Symmetrical

   B. Negatively skewed

   C. Positively skewed

   D. Leptokurtic

2. A distribution that is **negatively skewed** will have:

   A. A longer tail on the left

B. A higher mean than median

C. A flat peak

D. No tail at all

3. A **leptokurtic** distribution has:

   A. A flat peak and light tails

   B. A sharp peak and heavy tails

   C. A symmetric shape and flat tails

   D. No skewness

4. What does it mean when a distribution has **heavy tails** ?

   A. More values near the mean

   B. More extreme values in both directions

   C. No values at the extremes

   D. Skewness is zero

5. A **platykurtic** distribution is:

   A. Flat-topped and has lighter tails

   B. Tall-peaked with outliers

   C. Always positively skewed

   D. Symmetrical and normal

Short Answer questions

1. What is **skewness** ? How can it help describe a dataset?
2. What is the difference between **positive skew** and **negative skew** ?
3. What is **kurtosis** ? How does it relate to the shape of a distribution?
4. What do we mean by **"heavy tails"** in a distribution? Use illustrations to explain this.
5. Why is understanding skewness and kurtosis important when analyzing data?

Part C: Application – Interpretation (3 points each)

1. A dataset has a **mean much greater than the median** . What kind of skewness is likely present? Explain.
2. A distribution shows a **very tall peak** and **many outliers in both tails** . What is the likely kurtosis? What does this suggest about the data?
3. You are analyzing two datasets:

   - Dataset A has a kurtosis of 2.
   - Dataset B has a kurtosis of 6.

   Which has heavier tails? What does this mean in practice?

Percentiles

Quiz

1. The 90th percentile means:
    A. 90% of the values are above this point
    B. 90% of the values are below this point
    C. It is the average value
    D. It is the middle value
2. Which of the following is **always true** about the 50th percentile?
    A. It equals the mode
    B. It equals the mean
    C. It equals the median
    D. It equals the range
3. The 25th percentile is also known as:
    A. Q3
    B. Q1
    C. Median
    D. Mode
4. What is used to identify outliers using percentiles?
    A. Mean
    B. Standard deviation
    C. Interquartile range (IQR)
    D. Mode
5. Which of the following best describes an **outlier**?
    A. A value equal to the median
    B. A value far from the rest of the data
    C. A value between Q1 and Q3
    D. A missing value in the dataset

Short Answer Questions

1. What is a percentile? How is it used to describe a data point's position in a dataset?
2. How are percentiles and the interquartile range (IQR) used to identify outliers?
3. Explain how the 25th and 75th percentiles relate to the IQR.
4. Why might we remove or keep outliers in data analysis?
5. What do we mean when we say a value is "below the 5th percentile" or "above the 95th percentile"?

Numerical Practice

1. Find Q1, Q3, and IQR for the following data set: 10, 12, 13, 15, 18, 20, 22, 23, 25, 28 Then identify any potential outliers using the rule: Outliers are below Q1 - 1.5×IQR or above Q3 + 1.5×IQR.

2. A student scored at the 92nd percentile on a national exam. What does that tell you about their performance compared to others?

3. In a sample of ages: 21, 22, 23, 23, 25, 26, 29, 31, 80 Identify whether 80 is an outlier using the IQR method.