assignment=27

July 28, 2023

1 Q1. What is Statistics?

Statistics is a branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of data. It involves the use of quantitative methods to describe, summarize, and draw conclusions from data. Statistics is used in a wide variety of fields, including science, engineering, medicine, business, social sciences, and many others.

There are two main branches of statistics: descriptive statistics and inferential statistics. Descriptive statistics involves the use of methods to summarize and describe data, such as measures of central tendency (mean, median, mode), measures of variability (range, variance, standard deviation), and graphical displays (histograms, scatterplots, etc.). Inferential statistics involves the use of methods to make inferences or predictions about a population based on a sample of data. This includes hypothesis testing, estimation, and regression analysis.

Statistics plays a vital role in decision-making in many fields, as it allows us to analyze and interpret data to make informed decisions.

2 Q2. Define the different types of statistics and give an example of when each type might be used.

There are two main types of statistics: descriptive statistics and inferential statistics.

Descriptive Statistics: Descriptive statistics involves the use of methods to summarize and describe data. Some examples of descriptive statistics include: Measures of central tendency: Mean, median, mode. Measures of variability: Range, variance, standard deviation. Graphical displays: Histograms, scatterplots, boxplots. Descriptive statistics are useful when you want to summarize or describe a set of data. For example, if you want to describe the average income of a group of people, you could use the mean as a measure of central tendency. If you want to describe how spread out the incomes are, you could use the standard deviation as a measure of variability.

Inferential Statistics: Inferential statistics involves the use of methods to make inferences or predictions about a population based on a sample of data. Some examples of inferential statistics include: Hypothesis testing: Used to determine if there is a significant difference between two groups or if a particular relationship exists between two variables. Estimation: Used to estimate population parameters from sample data. Regression analysis: Used to model the relationship between two or more variables. Inferential statistics are useful when you want to make predictions or inferences about a larger population based on a smaller sample of data. For example, if you want to determine if a new drug is effective, you could use hypothesis testing to compare the outcomes of a treatment

group and a control group. If you want to estimate the proportion of a population that supports a particular policy, you could use estimation based on a sample of survey data.

3 Q3. What are the different types of data and how do they differ from each other? Provide an example of each type of data.

Data is defined as a systematic record corresponding to a specific quantity. Basically, data can be summarised as a set of facts and figures which can be used to serve a specific usage or purpose. For instance, data can be used as a survey or an analysis. Data in a systematic and organized form is referred to as information. In addition to this, the source of data primary or secondary is also an essential factor.

Types of Data In Statistics: In statistics, there are four main types of data: nominal, ordinal, interval, and ratio. These types of data are used to describe the nature of the data being collected or analyzed, and they help determine the appropriate statistical tests to use. In this essay, we will explore each type of data in detail, providing examples along the way.

Nominal Data: Nominal data is a type of data that consists of categories or names that cannot be ordered or ranked. Nominal data is often used to categorize observations into groups, and the groups are not comparable. In other words, nominal data has no inherent order or ranking. Examples of nominal data include gender (male/female), race (White/Black/Asian), religion (Christianity/Islam/Judaism), and blood type (A/B/AB/O).

Nominal data can be represented using frequency tables and bar charts, which display the number or proportion of observations in each category. For example, a frequency table for gender might show the number of males and females in a sample of people. A bar chart might display the proportions of males and females in the sample.

Nominal data is analyzed using non-parametric tests, which do not make any assumptions about the underlying distribution of the data. Common non-parametric tests for nominal data include chi-squared tests and Fisher's exact tests. These tests are used to compare the frequency or proportion of observations in different categories.

Ordinal Data: Ordinal data is a type of data that consists of categories that can be ordered or ranked. However, the distance between categories is not necessarily equal. Ordinal data is often used to measure subjective attributes or opinions, where there is a natural order to the responses. Examples of ordinal data include education level (elementary/middle/high school/college), job position (manager/supervisor/employee), and Likert scales (strongly agree/agree/disagree/strongly disagree).

Ordinal data can be represented using frequency tables, bar charts, or line charts. These displays show the order or ranking of the categories, but they do not imply that the distances between categories are equal.

Ordinal data is analyzed using non-parametric tests, which make no assumptions about the underlying distribution of the data. Common non-parametric tests for ordinal data include the Wilcoxon signed-rank test and the Mann-Whitney U test. These tests are used to compare the median or rank of observations in different categories.

Interval Data: Interval data is a type of data that consists of numerical values where the distance between each value is equal. However, there is no true zero point. Interval data is often used

to measure attributes such as temperature, dates, and time. Examples of interval data include temperature (Celsius/Fahrenheit), dates (days/months/years), and time (hours/minutes/seconds).

Interval data can be represented using histograms, boxplots, or line charts. These displays show the range of the data and the frequency or proportion of observations at each value.

Interval data is analyzed using parametric tests, which assume that the underlying distribution of the data is normal or approximately normal. Common parametric tests for interval data include the t-test, ANOVA, and regression analysis. These tests are used to compare the means or variances of observations in different groups or to examine the relationship between variables.

Ratio Data: Ratio data is a type of data that has a true zero point and an equal distance between each value. Ratio data is considered the most informative type of data because it can be used to make meaningful comparisons and calculations. In addition, ratio data can be used to perform all types of statistical analyses.

Examples of ratio data include height (inches/centimeters), weight (pounds/kilograms), income (dollars), and distance (miles/kilometers). For instance, if someone's height is 60 inches, it means that they are 5 feet tall, and if their height is 72 inches, it means that they are 6 feet tall. Moreover, if someone's weight is 150 pounds, it means that they weigh 68 kilograms.

Ratio data can be represented using histograms, boxplots, or line charts. These displays show the range of the data and the frequency or proportion of observations at each value. In addition, ratio data can be used to calculate various measures of central tendency, such as the mean, median, and mode, and measures of variability, such as range, variance, and standard deviation.

Ratio data is analyzed using parametric tests, which assume that the underlying distribution of the data is normal or approximately normal. Common parametric tests for ratio data include the t-test, ANOVA, and regression analysis. These tests are used to compare the means or variances of observations in different groups or to examine the relationship between variables.

Question 1. Difference between Quantitative data and Qualitative data?

Solution:

Quantitative data

Qualitative data

Data is depicted in numerical terms. Data is not depicted in numerical terms. Can be shown in numbers and variables like ratio, percentage, and more. Could be about the behavioral attributes of a person, or thing. Example: 100%, 1:3, 123 Example: loud behavior, fair skin, soft quality, and more. Question 2. Difference between Discrete and Continuous Data?

Solution:

Discrete Data

Continuous Data

The type of data that has clear spaces between values is discrete data. This information falls into a continuous series. Countable. Measurable There are distinct or different values in discrete data. Every value within a range is included in continuous data. Depicted using bar graphs Depicted using histograms Ungrouped frequency distribution of discrete data is performed against a single

value. Grouped distribution of continuous data tabulation frequencies is performed against a value group. Question 3. Give any two examples of data collection.

Solution:

Increase in population of our country in the last two decades. Number of rupees in the bag Question 4. Illustrate:

A. Describe how was your overall experience using the product?

B. Describe how was your overall experience using the product?

Good Poor What type of data is illustrated by these points.

Solution:

A reflects nominal data whereas B reflects ordinal data.

4 Q4. Categorise the following datasets with respect to quantitative and qualitative data types:

```
(i) Grading in exam: A+, A, B+, B, C+, C, D, E
 (ii) Colour of mangoes: yellow, green, orange, red
 (iii) Height data of a class: [178.9, 179, 179.5, 176, 177.2, 178.3, 175.8,...]
  (iv) Number of mangoes exported by a farm: [500, 600, 478, 672, ...]
```

```
(i) Grading in exam: A+, A, B+, B, C+, C, D, E - Qualitative data
(ii) Colour of mangoes: yellow, green, orange, red - Qualitative data
 (iii) Height data of a class: [178.9, 179, 179.5, 176, 177.2, 178.3, 175.8,...] - Quantitative
 (iv) Number of mangoes exported by a farm: [500, 600, 478, 672, ...] - Quantitative data
```

5 Q5. Explain the concept of levels of measurement and give an example of a variable for each level.

Levels of measurement, also known as scales of measurement, refer to the ways in which variables can be classified based on the nature of the data they represent. There are four commonly recognized levels of measurement: nominal, ordinal, interval, and ratio.

Nominal level of measurement: This level of measurement is used to classify data into categories that do not have any order or rank. The categories are simply used to label or identify items. Examples of variables at the nominal level of measurement include gender (male, female), eye color (brown, blue, green), and race (Caucasian, African American, Asian).

Ordinal level of measurement: This level of measurement allows variables to be ranked or ordered based on some criterion or characteristic. However, the differences between the categories are not necessarily equal or measurable. Examples of variables at the ordinal level of measurement include education level (high school diploma, bachelor's degree, master's degree), socioeconomic status (low, middle, high), and grade in a course (A, B, C, D, F).

Interval level of measurement: This level of measurement is similar to the ordinal level, but the differences between the categories are equal and measurable. There is no true zero point in interval

data. Examples of variables at the interval level of measurement include temperature (Fahrenheit or Celsius), IQ scores, and dates (measured in years or days from a certain point).

Ratio level of measurement: This level of measurement is the most precise and has a true zero point. Ratio variables can be ordered, and the differences between categories are both equal and meaningful. Examples of variables at the ratio level of measurement include height, weight, time, and income.

Overall, understanding the level of measurement of a variable is important because it affects the type of statistical analysis that can be performed on the data. For instance, nominal and ordinal data may only be analyzed using non-parametric statistical tests, while interval and ratio data can be analyzed using both non-parametric and parametric statistical tests.

6 Q6. Why is it important to understand the level of measurement when analyzing data? Provide an example to illustrate your answer.

Understanding the level of measurement is essential when analyzing data because it determines the type of statistical analysis that can be performed and the appropriate measures of central tendency and dispersion that can be used. Different levels of measurement have different properties, which affect the way data can be analyzed and interpreted.

For example, if we have data on the favorite color of a group of people, we need to know the level of measurement of the variable to perform appropriate analysis. If the variable is measured on a nominal scale, we can only count the number of people who choose each color and calculate the percentage of people for each color. We cannot compute a mean or variance because nominal data have no natural ordering or hierarchy. On the other hand, if the variable is measured on an ordinal scale, we can rank the colors from most to least popular and compute the median or mode. However, we cannot assume that the intervals between ranks are equal, and we cannot compute a meaningful mean or variance.

In contrast, if the variable is measured on an interval or ratio scale, we can assume equal intervals between values and compute a meaningful mean and variance. For example, if we have data on the age of a group of people, we can calculate the mean, median, and standard deviation, which provide more detailed information about the distribution of the data.

In summary, understanding the level of measurement is crucial because it determines the appropriate statistical analysis and measures of central tendency and dispersion that can be used. It helps researchers avoid making incorrect assumptions about the properties of the data and make more accurate conclusions based on the analysis.

7 Q7. How nominal data type is different from ordinal data type.

Nominal and ordinal are two of the four levels of measurement in statistics.

Nominal data is a type of categorical data in which data values represent categories that do not have any inherent order or ranking. For example, gender (male, female) or color (red, green, blue) are nominal variables. Nominal data can only be classified and counted.

Ordinal data, on the other hand, represents categories that can be ranked in order or have a natural order. However, the distance between categories is not equal. For example, the ranking of the finishing positions in a race (1st, 2nd, 3rd) is an ordinal variable. Other examples of ordinal variables include education level (high school, college, graduate degree) or ratings (poor, fair, good, excellent).

The key difference between nominal and ordinal data is that nominal data does not have any inherent order or ranking, whereas ordinal data has a natural order or ranking but no specific numerical value associated with it.

8 Q8. Which type of plot can be used to display data in terms of range?

A box plot, also known as a box-and-whisker plot, is commonly used to display data in terms of range. It shows the distribution of the data by displaying the median, quartiles, and outliers of the dataset. The box represents the interquartile range (IQR), which is the range of the middle 50% of the data, while the whiskers extend to the minimum and maximum values within a certain range. Box plots are useful for comparing distributions between different groups or variables.

9 Q9. Describe the difference between descriptive and inferential statistics. Give an example of each type of statistics and explain how they are used.

Descriptive statistics and inferential statistics are two broad categories of statistical analysis methods used in data analysis.

Descriptive statistics is the branch of statistics that deals with the summary and description of the characteristics of a set of data. It includes measures like mean, median, mode, standard deviation, range, and percentiles, which are used to describe and summarize data in a way that is easily understandable. For example, if we want to understand the average income of a particular community, we can calculate the mean income of that community using descriptive statistics.

In contrast, inferential statistics is the branch of statistics that deals with making generalizations and predictions about a population based on a sample. It involves using sample data to make inferences or draw conclusions about a larger population. For example, if we want to know whether the average income of a particular community is significantly different from the average income of the entire population, we can use inferential statistics to conduct hypothesis testing and make conclusions about the population based on the sample data.

To illustrate the difference between descriptive and inferential statistics, consider the following example. Suppose we want to study the average height of students in a school. Descriptive statistics can be used to calculate the mean, median, and range of the heights of the students in the school. On the other hand, inferential statistics can be used to make generalizations about the height of all students in the school, based on a sample of students. For instance, we can randomly select a sample of 100 students, calculate their average height, and use inferential statistics to make predictions about the average height of all students in the school.

10 Q10. What are some common measures of central tendency and variability used in statistics? Explain how each measure can be used to describe a dataset.

Measures of central tendency and variability are commonly used in statistics to describe a dataset.

Measures of central tendency:

Mean: The mean is the sum of all the values in a dataset divided by the number of values in the dataset. It is often used to describe the average value of a dataset. For example, the mean salary of employees in a company can be used to describe the average salary of all employees.

Median: The median is the middle value of a dataset when it is ordered from smallest to largest. It is often used to describe the central value of a dataset when there are outliers or when the distribution is not symmetrical. For example, the median income of a neighborhood can be used to describe the central income of the residents.

Mode: The mode is the value that appears most frequently in a dataset. It is often used to describe the most common value of a dataset. For example, the mode of transportation used by commuters in a city can be used to describe the most common method of transportation.

Measures of variability:

Range: The range is the difference between the largest and smallest values in a dataset. It is often used to describe the spread of the data. For example, the range of ages of students in a classroom can be used to describe how spread out the ages are.

Variance: The variance is the average of the squared differences from the mean. It is often used to describe how spread out the data is from the mean. For example, the variance of test scores in a class can be used to describe how spread out the scores are from the average score.

Standard deviation: The standard deviation is the square root of the variance. It is often used to describe how spread out the data is from the mean in the same units as the data. For example, the standard deviation of the heights of students in a class can be used to describe how spread out the heights are from the average height, in units of centimeters.

These measures of central tendency and variability can be used together to provide a more complete description of a dataset. For example, the mean and standard deviation can be used to describe the average value and the spread of the data, while the median and range can be used to describe the central value and the range of the data.

[ ]:

[ ]:

[ ]: