DIVING INTO TO THE WORLD OF DATA (2022) INTERVIEW QUESTIONS

Search...

1. Why would you use the Median as a measure of central tendency?

The median is a statistical measure of central tendency that is used in certain situations when analyzing data. Here are some reasons why the median might be preferred:

Skewed data: The median is less affected by outliers or extreme values compared to other measures of central tendency like the mean. If the data has a skewed distribution or contains outliers, the median provides a more robust representation of the center of the data.

Ordinal data: The median is particularly useful when dealing with ordinal data, where the values can be ranked but do not have a precise numerical meaning. In such cases, taking the average (mean) of the data wouldn't make sense, but finding the middle value (median) can still provide meaningful information about the central tendency.

Non-normal distributions: If the data does not follow a normal distribution, using the median can be beneficial. The mean is heavily influenced by the shape of the distribution, and in non-normal distributions, it may not accurately represent the center of the data. The median, on the other hand, does not rely on the specific shape of the distribution and can provide a more representative measure.

As an alternative perspective: Sometimes, using the median alongside other measures of central tendency like the mean can provide a more comprehensive understanding of the data. While the mean represents the arithmetic average and the sum of all values, the median offers a different perspective by focusing on the middle value. By considering both measures, you can gain a more complete picture of the central tendency.

It's important to note that the choice of using the median as a measure of central tendency depends on the nature of the data and the specific goals of the analysis.

2. What is Central Tendency?

Central tendency is a statistical concept that refers to the measure or value that represents the center or typical value of a dataset. It provides a summary of the location or concentration of the data points.

There are several common measures of central tendency, including:

Mean: The mean, often referred to as the average, is calculated by summing up all the values in the dataset and dividing by the total number of values. It is commonly used when the data follows a normal distribution or when the values are numerical and have a meaningful quantitative interpretation.

Median: The median is the middle value in an ordered dataset. To find the median, the data points are arranged in ascending or descending order, and the middle value is selected. If the dataset has an even number of values, the median is calculated as the average of the two middle values. The median is useful when dealing with skewed data or when the values are ordinal and lack a precise numerical meaning.

Mode: The mode represents the value(s) that occur most frequently in the dataset. Unlike the mean and median, which rely on numerical calculations, the mode is based on the count of values. The mode is particularly useful when dealing with categorical or nominal data.

These measures of central tendency provide different perspectives on the center of the data and are applicable in different scenarios. The choice of measure depends on the nature of the data, its distribution, and the specific objectives of the analysis.

3. What is the difference between Descriptive Statistics and Inferential Statistics?

Descriptive statistics and inferential statistics are two branches of statistics that serve different purposes in data analysis. Here's an explanation of their differences:

Descriptive Statistics:

Descriptive statistics involve methods and techniques used to summarize and describe the main features or characteristics of a dataset. It focuses on organizing, presenting, and summarizing the data in a meaningful way. Descriptive statistics provide a way to understand and interpret the data by using measures such as central tendency (mean, median, mode), measures of dispersion (variance, standard deviation, range), and graphical representations (histograms, bar charts, pie charts). The goal of descriptive statistics is to provide a concise summary of the data and gain insights about its characteristics.

Inferential Statistics:

Inferential statistics involves making inferences, predictions, or generalizations about a population based on a sample of data. It uses probability theory and statistical techniques to draw conclusions beyond the observed data. Inferential statistics allows researchers to make claims or hypotheses about a larger population based on the analysis of a smaller sample. It involves estimating parameters, testing hypotheses, determining relationships between variables, and making predictions. By using inferential statistics, researchers can draw conclusions about populations and generalize their findings beyond the specific dataset they have analyzed.

In summary, descriptive statistics focuses on summarizing and describing the data at hand, while inferential statistics aims to make broader inferences and generalizations about populations based on sample data. Descriptive statistics provides insights into the characteristics of the dataset, while inferential statistics extends those insights to a larger context and helps draw conclusions about the population from which the data is sampled.

4. Can there be more than one Mode?

Yes, it is possible for a dataset to have more than one mode. In statistics, the mode refers to the value(s) that occur most frequently in a dataset. If multiple values have the same highest frequency and occur more frequently than any other value in the dataset, then there are multiple modes.

When a dataset has a single mode, it is called unimodal. For example, in a dataset {2, 4, 4, 6, 8}, the mode is 4 because it occurs more frequently than any other value.

However, there are cases where a dataset may have more than one mode. When a dataset has two modes, it is called bimodal. For example, in a dataset {2, 4, 4, 6, 6, 8}, both 4 and 6 occur with the same highest frequency, so there are two modes.

Similarly, a dataset can have more than two modes, which is referred to as multimodal. For example, in a dataset {2, 4, 4, 6, 6, 8, 8}, the values 4, 6, and 8 all occur with the same highest frequency, so there are three modes.

It's important to note that there can also be cases where a dataset has no mode if no value occurs more frequently than others or if all values occur with the same frequency.

So, the presence and number of modes in a dataset depend on the frequencies of the values and whether there is a tie for the highest frequency.

5. In which use-case we should use Mean and when to use Median?

The choice between using the mean or median as a measure of central tendency depends on the nature of the data and the specific goals of the analysis. Here are some guidelines:

Use Mean:

When the data is approximately normally distributed: The mean is suitable when the data follows a symmetric bell-shaped distribution. In such cases, the mean provides a precise measure of central tendency.

When the data is numerical and has a meaningful quantitative interpretation: The mean is appropriate for data with a numeric scale where the values represent quantities. For example, the mean is commonly used for variables such as age, height, weight, or test scores.

When outliers are not present or are not of concern: The mean is sensitive to extreme values or outliers. If the dataset contains outliers that could heavily influence the central tendency, it might be better to use other measures like the median.

Use Median:

When the data is skewed or has outliers: The median is more robust to extreme values and skewed distributions. It represents the middle value in the dataset and is less influenced by outliers, making it a suitable choice when the data deviates from a normal distribution.

When dealing with ordinal or ranked data: The median is appropriate for data that can be ordered but does not have a precise numerical meaning. For example, when analyzing survey responses using Likert scales or rankings, the median can provide a meaningful measure of central tendency.

When the data is discrete or categorical: The median can be useful when working with discrete or categorical variables where the concept of an average may not be applicable. In such cases, finding the middle value can provide a representative measure.

In some situations, using both the mean and median together can provide a more comprehensive understanding of the data. By considering both measures, you can gain insights into different aspects of the central tendency and the distribution of the data.

6. What does a Statistical Test do?

A statistical test is a formal method used to make inferences or draw conclusions about a population based on sample data. It allows researchers to assess whether observed differences or relationships in the sample are statistically significant, meaning they are unlikely to have occurred by chance.

Statistical tests involve the following general steps:

Formulating hypotheses: The first step is to formulate a null hypothesis (H0) and an alternative hypothesis (Ha). The null hypothesis typically represents the status quo or no effect, while the alternative hypothesis represents the presence of an effect or a difference.

Choosing a test statistic: The test statistic is a numerical value calculated from the sample data. It quantifies the degree of difference or relationship between variables being tested. The choice of the test statistic depends on the specific research question and the nature of the data.

Determining the significance level: The significance level, denoted as α (alpha), is the threshold chosen to assess the strength of evidence against the null hypothesis. It represents the probability of erroneously rejecting the null hypothesis when it is actually true. Commonly used significance levels are 0.05 (5%) or 0.01 (1%).

Conducting the test and calculating the p-value: The test is performed by calculating the test statistic using the sample data. The p-value is then calculated, which represents the probability of obtaining a test statistic as extreme or more extreme than the one observed, assuming the null hypothesis is true.

Interpreting the results: The p-value is compared to the significance level to make a decision regarding the null hypothesis. If the p-value is less than the significance level, typically α, the null hypothesis is rejected in favor of the alternative hypothesis, indicating a statistically significant result. If the p-value is greater than the significance level, there is insufficient evidence to reject the null hypothesis.

Statistical tests help researchers make informed decisions based on evidence from the data. They provide a way to determine if observed differences or relationships are likely to be due to chance or if they represent true effects or associations in the population.

7. When would you use the Interquartile Range (IQR)?

The interquartile range (IQR) is a statistical measure that describes the spread or dispersion of a dataset. It is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). The IQR is useful in several scenarios:

Identifying outliers: The IQR is commonly used to detect outliers in a dataset. Outliers are values that fall significantly below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR. By examining values outside this range, you can identify potential extreme observations that may warrant further investigation.

Describing the spread of the middle 50% of data: The IQR provides a measure of dispersion for the middle 50% of the data. It focuses on the range where the majority of observations lie, excluding outliers. The IQR is less influenced by extreme values than the range, making it a useful measure when the data contains outliers.

Comparing variability between groups: The IQR can be used to compare the variability of different groups or datasets. By comparing the IQRs of multiple groups, you can assess whether the spreads of the distributions differ. If the IQRs are similar, it suggests comparable variability between the groups, while a larger IQR indicates greater variability.

Summarizing skewed or non-normal data: The IQR is particularly helpful when dealing with skewed or non-normal distributions. Unlike the mean and standard deviation, the IQR is not affected by extreme values or the shape of the distribution. It provides a robust measure of dispersion that is less influenced by outliers or deviations from normality.

In summary, the interquartile range (IQR) is used to understand the spread of the central portion of a dataset, identify outliers, compare variability between groups, and summarize data that deviates from a normal distribution. It is a versatile measure that is less sensitive to extreme values, making it suitable for a variety of data analysis situations.

8. How many types of measures of Variability do you know?

There are several common measures of variability used in statistics to quantify the spread, dispersion, or variability of a dataset. Here are some of the main measures:

Range: The range is the simplest measure of variability and is calculated as the difference between the maximum and minimum values in a dataset. While easy to compute, it is highly influenced by outliers and may not provide a comprehensive understanding of the spread.

Interquartile Range (IQR): The interquartile range is calculated as the difference between the third quartile (Q3) and the first quartile (Q1) in a dataset. It represents the spread of the middle 50% of the data and is less affected by outliers compared to

the range.

Variance: Variance is a measure that quantifies the average squared deviation of each data point from the mean. It considers all values in the dataset and provides an indication of the overall variability. However, it is influenced by extreme values due to the squaring of deviations.

Standard Deviation: The standard deviation is the square root of the variance. It is a commonly used measure that represents the average deviation from the mean. The standard deviation is widely used due to its intuitive interpretation and compatibility with the normal distribution.

Mean Absolute Deviation (MAD): The mean absolute deviation calculates the average absolute difference between each data point and the mean. It is less influenced by extreme values compared to variance and standard deviation because it does not involve squaring the differences.

Coefficient of Variation (CV): The coefficient of variation is a relative measure of variability expressed as a percentage. It is calculated by dividing the standard deviation by the mean and multiplying by 100. The CV is useful for comparing the variability of different datasets with different units or scales.

These are some of the commonly used measures of variability. The choice of which measure to use depends on the specific characteristics of the dataset, the goals of the analysis, and the presence of outliers or extreme values.

9. What are the different types of data?

In statistics, data can be classified into different types based on their characteristics and properties. The main types of data are:

Nominal Data: Nominal data consists of categories or labels that do not have a specific order or numerical meaning. Examples include gender (male/female), eye color (blue/brown/green), or types of fruits (apple/orange/banana). Nominal data can be represented using words, letters, or symbols.

Ordinal Data: Ordinal data represents categories with a specific order or ranking. While the categories have a relative position, the numerical difference between them may not be meaningful or consistent. Examples of ordinal data include educational levels (high school/college/graduate), customer satisfaction ratings (low/medium/high), or survey responses using Likert scales (strongly disagree/disagree/neutral/agree/strongly agree).

Interval Data: Interval data is numerical data that has consistent intervals between values, but does not have a true zero point. The zero in interval data is arbitrary and does not indicate the absence of the measured quantity. Examples include temperatures in Celsius or Fahrenheit, years on the calendar, or IQ scores. Arithmetic operations like addition and subtraction can be performed on interval data.

Ratio Data: Ratio data is similar to interval data, but with a true zero point. The zero in ratio data represents the absence of the measured quantity. Examples of ratio data include height, weight, time, distance, or counts of occurrences. All arithmetic operations (addition, subtraction, multiplication, division) are meaningful for ratio data.

Continuous Data: Continuous data is numerical data that can take on any value within a range. It is often measured using instruments or devices. Examples of continuous data include height, weight, temperature, or time. Continuous data can be represented as real numbers and allows for infinite possible values within a given range.

Discrete Data: Discrete data consists of separate, distinct values or counts that are typically whole numbers. Discrete data represents data that can only take specific values and cannot be subdivided further. Examples include the number of children in a family, the number of cars in a parking lot, or the number of items sold.

Understanding the type of data is important as it determines the appropriate statistical analysis methods and techniques to be used. Different types of data require different approaches for analysis, summarization, and interpretation.

10. Which visualization should be used for the different types of data?

The choice of visualization depends on the type of data and the specific goals of the analysis. Here are some commonly used visualizations for different types of data:

Nominal Data:

Bar Chart: A bar chart is suitable for displaying the frequency or proportion of different categories. Each category is represented by a bar, and the height of the bar corresponds to the frequency or proportion of that category.

Pie Chart: A pie chart is useful for illustrating the proportion of each category in relation to the whole. It is suitable when you want to visualize the distribution of categorical data.

Ordinal Data:

Bar Chart: Similar to nominal data, a bar chart can be used to display the frequency or proportion of each ordinal category. The categories are arranged in a specific order to reflect their ranking or relative position.

Interval/Ratio Data:

Histogram: A histogram is a graphical representation of the distribution of numerical data. It displays the frequency or density of data within specified intervals or bins. Histograms are useful for visualizing the shape, central tendency, and spread of continuous data.

Box Plot: A box plot (or box-and-whisker plot) is used to display the distribution of numerical data through quartiles. It provides information about the median, quartiles, and potential outliers, making it useful for comparing groups or visualizing the spread of data.

Line Chart: A line chart is often used to show the trend or change in numerical data over time or another continuous variable. It is suitable for visualizing the relationship between two variables and identifying patterns or trends.

Discrete Data:

Bar Chart: Similar to nominal data, a bar chart is useful for displaying the frequency or count of discrete data categories. Each category is represented by a bar, and the height of the bar corresponds to the frequency or count.

Line Chart: A line chart can be used to show the trend or change in discrete data over time or another continuous variable. It can be useful for illustrating patterns or changes in counts or frequencies.

Remember, these are general guidelines, and the choice of visualization can vary depending on the specific context, data characteristics, and research questions. It's important to select a visualization that effectively communicates the information and insights you want to convey.