Statistics and Probability Interview Questions

1. What is measure of central tendency? Give examples: Also, where to use Mean for the measure of centre and where to use median for the measure of center. (with examples)

The measure of the central tendency is a statistical concept used to describe the center or typical value of the dataset. It provides the single value that represents the entire dataset.

There are three commonly used measures of central tendancy:

1.Mean: The mean is calculated by summing all the values in a dataset by dividing it by the total number of values. It is suitable dataset with a symmetric distribution or when the datapoints are evenly distributed. For example, if we have the following dataset of ages 20,25,30,35 and 40. The mean age would be (20+25+30+35+40)/5 = 30

2.Median : The median is the middle value of a dataset when it is arranged in ascending or descending order. It is suitable for datasets with skewed distibutions or when there are outluers present. For example, if we have the following dataset of incomes : 25000,30000,35000,40000,100000, the median income would be 35000.

3.Mode : The mode is the most frequently occurring value in the dataset. It is suitable for the dataset with the categorical or discrete variables. For examples, if we have the following dataset of eye colors: blue,green, brown, blue , the mode eye color would be blue.

The choice between using mean or median as a measure of central tendancy depends on:

* Distribution and charecteristics of the data.
* Mean is sensitive to extreme values or outliers and is appropriate when the data is normally distributed.
* The median is more robust to outliers and is suitable when the the data is skewed or contains extreme values.

1. **What is the measure of spread? What is the difference between the standard deviation and Interquartile range? When to use that?**

The Measure Of Spread, also known as variability or dispersion, is a statistical concept that quanitifies the extent to which the data points are spread out or clustered.

It provides information about the spread or dispersion of data values around the central tendancy.

Std. dev. And IQR is most commonly used measure of spreads.

**Standard deviation** is a measure of average distance between each data point and the Mean of the data set. It takes into account all the data points and provides a measure of overall variability in the data. It is calculated by taking the square root of the variance. The standard deviation is sensitive to outliers and extreme values in the dataset.

**IQR** on the other hand, is a measure of the spread of the middle 50% of the data. It is calculated. It is calculated as the difference between the third quartile and first quartile of the dataset. The interquartile range is less affected by outliers and extreme values as it considered the range of values of the middle 50% of the dataset.

When to use the standard deviation and IQR is depending upon the nature of the data and the specific analysis or the context.

* Standard deviation is commonly used when the data follows a normal distribution with no extreme outliers. It provides a comprehensive measure of spread that considers all data points. It is also used in many statistical and hypothesis testing.
* IQR is useful when dataset has outliers or the extreme values that could significantly affect the standard deviation. It provides the measure of spread of the central position of the data, which can be more robust to outliers. IQR often used in EDA and when comparing the spread datasets with different distributions or outliers.
* In summary standard deviation is more sensitive to outliers. It provides the measure of overall data spread, while the interquartile range is less affected by outliers and focuses on the spread of the middle 50% of the data. The choice between the two depends on the specific characteristics of the data and the objectives of the analysis.

1. What is the definition of normal distribution? How do you calculate the Mean and standard deviation of a normal distribution? What is the empirical rule (68-95-99.7), and how does it relate to the normal distribution? How do you standardize a variable using z-score within a normal distribution? What is the relationship between the normal distribution and standard normal distribution?

A normal distribution is also known as guassian distribution, is a probability distribution that is symmetric and bell-shaped. It is characterized by its mean and standard deviation.

In normal distribution mean represents the central tendency of the data.

While standard deviation measures the spread or dispersion of the data.

To calculate mean of normal distribution, you sum up all the values in the data set and divide it by the total number of values. Mean = (x1+x2+x3…..+xn)/n where x1,x2,x3….xn are the individual values and n is the total number of values.

To calculate standard deviation of a normal distribution, first calculate the variance by finding the average squared difference between each data point and mean.

Variance = [(x1-mean)^2+(x2-mean)^2…..+(xn+mean)^2]/n

If you take the square root of the variance to get standard deviation.

* The impirical rule, also known as the 68-95-99.7 rule, is a rule of thumb that applies to normal distributions. It states that approximately **68%** of the data falls within one standard deviation, about **95%** falls within two standard deviation and nearly **99.7%** of the data falls within three standard deviations.
* To standardize a variable using **z-score** within normal distribution, you subtract the Mean from the individual value and divide it by the standard deviation. The formula for calculating the z-score is

Z= (x=u)/std.dev. , where x is the individual value, u is the mean and standard deviation.

This process transforms the variable into a **standard normal distribution** , a z-distribution, with a mean of 0 and standard deviation as 1.

* The relationship between the **normal** and the **standard** **normal** **distribution** is that any normal distribution can be transformed into a **standard normal distribution** using the process of **standardization**.
* This allows for **easier comparison** and **analysis** of different normal distributions. The standard normal distributions serves as a reference distributions, with its properties well-known and extensively studied.

1. Explain the Central Limit Theorem. (Give practical example)

The distribution of the sample means will be normally distributed even if the original distribution is not normally distributed. (regardless of the shape of the original population distribution). In simpler terms, it suggests that as the sample size increases,the distribution of sample means will become more and more like normal distribution.

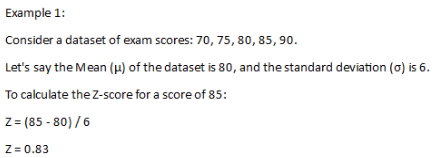
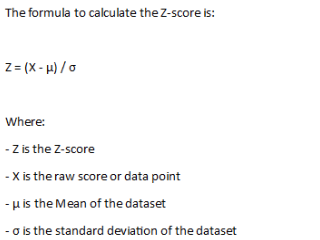
Practical examples of the central limit theorem.

1. Height of individuals: If you measure the heights of a large number of people in population, the individual heights may not follow a normal distribution. However, if you take random samples of a certain size (e.g.30) and calculate the mean height for each sample. The distribution of these samples means will be approximately normally distributed.
2. Test scores: Consider a class of students whose test scores are not normally distributed. If you randomly select samples of a specific size (e.g. 50) from this class and calculate the mean test score for each of the sample, the distribution of these samples means will become more and more normal as the sample size increases.
3. Coin flip: Suppose you repeatedly flip a biased coin,where the probability of getting heads is not 0.5. if you take multiple samples of certain size(e.g. 100) and calculate the proportion of the heads in each sample. These sample proportions will approach a normal distribution.

The central limit theorem is crucial in statistics as it allows us to make inference about a population based on the sample. It enables us to use the properties of normal distributions to estimate the population parameters and make statistical decisions.

5. What is Z-score? How do we calculate the same?

Z-score is also known as standard score, is a statistical measurement that quantifies the number of standard deviations a data point is from Mean of the dataset. It is used to standardize and compare values from different datasets.



In example, the Z-score represents the number of standard deviations the given data point is away from the mean. A positive Z-score indicates that the data point is above the Mean, while negative z-score indicates that it is below the mean value.

6.