STATISTICS

Basically, the statistical analysis is meant to collect and study the information available in large quantities. Statistics is a branch of mathematics, where computation is done over a bulk of data using charts, tables, graphs, etc.

Def: The data collected for analysis here is called measurements. Now, if we have to measure the data based on a scenario, a sample is taken out of a population. Then the analysis or calculation is done for the following measuremen

1 What is Quantitaive and Qualitative Statistics? Example

Collected data can be statistically analyzed
Examples: Height, Weight, Time, Price, Temperature, etc.

Collected data can just be observed and not evaluated
Examples: Scents, Appearance, Beauty, Colors, Flavors, etc.

2 What is a Dichotomous Data type ? Example?

Dichotomous (outcome or variable) means "having only two possible values", e.g. "yes/no", "male/female", "head/tail", "age > 35 / age <= 35" etc.

Dichotomous variables are categorical variables with two levels. These could include yes/no, high/low, or male/female. To remember this, think di = two. Ordinal variables have two are more categories that can be ordered or ranked.

3 What is Estimate of Location?

Location estimation refers **to the process of obtaining location information on a node with respect to a set of known reference positions**.

Mean And Median are some of the possible estimates of location

4 What is the difference between Mean and Median? Example?

The mean (average) of a data set is found by adding all numbers in the data set and then dividing by the **number of values** in the set. The median is the middle value when a data set is ordered from least to greatest

Why do we use weighted mean?

Some values are intrinsically more variable than others, and highly variable observations are given a lower weight. For example, if we are taking the average from multiple sensors and one of the sensors is less accurate, then we might downweight the data from that sensor.

The data collected does not equally represent the different groups that we are interested in measuring. For example, because of the way an online experiment was conducted, we may not have a set of data that accurately reflects all groups in the user base. To correct that, we can give a higher weight to the values from the groups that were underrepresented.

5 What is a Normal Distribution ?

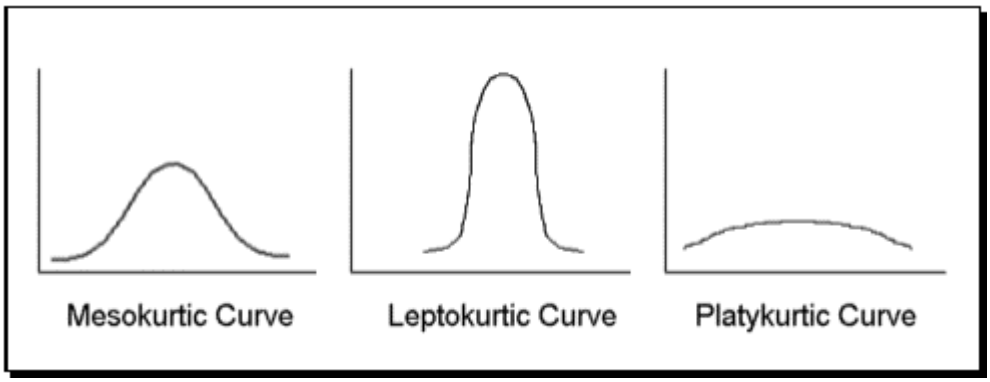
Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

The standard normal distribution, also called the z-distribution, is a special normal distribution where the mean is 0 and the standard deviation is 1. Any normal distribution can be converted into the standard normal distribution by turning the individual values into z-scores

A normal distribution is the proper term for a probability bell curve.
In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.

Kurtosis: describes the distribution of data around the average/mean.
It measures the peak if distributions.
mesokurtic - normal kurtosis of 3
Hight kurtosis has fatter tail and low has skinny tails and distribution concentrated moe towards the avg - leptokurtic.
low kurtosis has skinny/thinner tails with distrib evenly present with fewer values in its shorter (i.e. lighter and thinner) tails - palykurtic.

Mesokurtic: Distributions that are moderate in breadth and curves with a medium peaked height.
Leptokurtic: More values in the distribution tails and more values close to the mean (i.e. sharply peaked with heavy tails)
Platykurtic: Fewer values in the tails and fewer values close to the mean (i.e. the curve has a flat peak and has more dispersed scores with lighter tails).



6 What is Standard Deviation?

Formula

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

σ = population standard deviation
N = the size of the population
x_i = each value from the population
μ = the population mean

In statistics, the standard deviation is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range

For a normal distribution 68% values fall in first standard deviation(mue + sigma)
95% falls in second SD(mue + 2 sigma)
Note: 99.7% falls in 3rd SD

7 What is the difference between Standard Deviation and Variance?

variance is the measure of dispersion of data around mean

Variance is a measure of how data points vary from the mean, whereas standard deviation is the measure of the distribution of statistical data. The basic difference between both is standard deviation is represented in the same units as the mean of data, while the variance is represented in squared units.

Coefficient of variation - sigma/mue - SD/mean

8 What is MAD, MSE, RMSE?

Two of the most commonly used forecast error measures are mean absolute deviation (MAD) and mean squared error (MSE). MAD is the average of the absolute errors. MSE is the average of the squared errors. ... Either MAD or MSE can be used to compare the performance of different forecasting techniques.

The absolute error is the absolute value of the difference between the forecasted value and the actual value. MAE tells us how big of an error we can expect from the forecast on average. ... Mean Absolute Percentage Error (MAPE) allows us to compare forecasts of different series in different scales
The MSE is the sum of the squared errors divided by the number of observations. The Root Mean Square Error (RMSE) is the square root of the MSE. RMSE is used to convert MSE back into the same units as the actual data.

Difference between MAD and SD

Average absolute deviation ⋮ Overvie

Formula

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - m(X)|$$

$m(X)$ = average value of the data set
n = number of data values
x_i = data values in the set

The average deviation, or mean absolute deviation, is calculated similarly to standard deviation, but it uses absolute values instead of squares to circumvent the issue of negative differences between the data points and their means. To calculate the average deviation: Calculate the mean of all data points

While both measures rely on the deviations from the mean (x - \bar{x}), the MAD uses the absolute values of the deviations and the standard deviation uses the squares of the deviations. Both methods result in non-negative differences. The MAD is simply the mean of these nonnegative (absolute) deviations----------The average absolute deviation of a data set is the average of the absolute deviations from a central point. It is a summary statistic of statistical dispersion or variability

9 What is Hypothesis?

A statistical hypothesis is an assumption about a population parameter . This assumption may or may not be true. For instance, the statement that a population mean is equal to 10 is an example of a statistical hypothesis. A researcher might conduct a statistical experiment to test the validity of this hypothesis.

10 What is Inferential Statistics?

With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data alone. For instance, we use inferential statistics to try to infer from the sample data what the population might think. Or, we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study. Thus, we use inferential statistics to make inferences from our data to more general conditions; we use descriptive statistics simply to describe what's going on in our data.

11 What are the differnet types of statistics?

ypes of Statistics
Statistics is mainly divided into the following two categories.

Descriptive statistics
Inferential statistics

Descriptive Statistics
In the descriptive statistics, the data is described in a summarized way. The summarization is done from the sample of the population using different parameters like mean or standard deviation. Descriptive statistics are a way of using charts, graphs, and summary measures to organize, represent, and explain a set of data.
Data is typically arranged and displayed in tables or graphs summarizing details such as histograms, pie charts, bars or scatter plots.
Descriptive statistics are just descriptive and thus do not require generalization beyond the data collected.

Inferential Statistics
In the Inferential Statistics, we try to interpret the meaning of descriptive statistics. After the data has been collected, analyzed, and summarised we use Inferential Statistics to describe the meaning of the collected data.
Inferential Statistics use the probability principle to assess whether trends contained in the research sample can be generalized to the larger population from which the sample originally comes.
Inferential Statistics are intended to test hypotheses and investigate relationships between variables and can be used to make population predictions.
Inferential Statistics are used to draw conclusions and inferences, i.e., to make valid generalizations from samples.

Example
In a class, the data is the set of marks obtained by 50 students. Now when we take out the data average, the result is the average of 50 students' marks. If the average marks obtained by 50 students are 88 out of 100, on the basis of the outcome, we will draw a conclusion.

Stages Of Statistics

Stages of Statistics
Collection of Data:
This is the first step of statistical analysis where we collect the data using different methods depending upon the case.

Organizing the Collected Data:
In the next step, we organize the collected data in a meaningful manner. All the data is made easier to understand.

Presentation of Data:
In the third step we simplify the data. These data are presented in the form of tables, graphs, and diagrams.

Analysis of the Data:
Analysis is required to get the right results. It is often carried out using measures of central tendencies, measures of dispersion, correlation, regression, and interpolation.

Interpretation of Data:
In this last stage, conclusions are enacted. Use of comparisons is made. On this basis, forecasting is made.

Uses of Statistics
Statistics helps to obtain appropriate quantitative data.
Statistics helps to present complex data for the simple and consistent interpretation of the data in a suitable tabular, diagrammatic, and graphic form.
Statistics help to explain the nature and pattern of variability through quantitative observations of a phenomenon.
Statistics help to depict the data in tabular form, or in a graphical form in order to understand it properly.

Applications of Statistics
Statistics is used in Machine Learning and Data Mining.
Statistics is used in Mathematics.
Statistics is used in Economics.

12 What is Ztest? For what purpose it is used?

Z Test is the statistical hypothesis which is used in order to determine that whether the two samples means calculated are different in case the standard deviation is available and sample is large whereas the T test is used in order to determine a how averages of different data sets differs from each other in case standard deviation or the variance is not known
#1 – T-Test
Z-test Formula
, as mentioned earlier, are the statistical calculations that can be used to compare population averages to a sample's. The z-test will tell you how far, in standard deviations
terms, a data point is from the average of a data set. A z-test will compare a sample to a defined population that is typically used for dealing with problems relating to large samples (i.e., n > 30). Mostly, they are very useful when the standard deviation is known.

#2 – T-Test
T-tests
are also calculations that can be used to test a hypothesis, but they are very useful when we need to determine if there is a statistically significant comparison between the 2 independent sample groups. In other words, a t-test asks whether the comparison between the averages of 2 groups is unlikely to have occurred due to random chance. Usually, t-tests are more appropriate when dealing with problems with a limited sample size
(i.e., n < 30).

13 What is correlation and covariance?

Both covariance and correlation measure the relationship and the dependency between two variables. Covariance indicates the direction of the linear relationship between variables. Correlation measures both the strength and direction of the linear relationship between two variables

14 What is Normal Distribution

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

KEY TAKEAWAYS
A normal distribution is the proper term for a probability bell curve.
In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
Normal distributions are symmetrical, but not all symmetrical distributions are normal.
In reality, most pricing distributions are not perfectly normal

15 What is skewed Distribution & uniform distribution?

Skewed distribution is a condition when one side (either right or left) of the graph has more dataset in comparison to the other side. Uniform distribution is **a condition when all the observations in a dataset are equally spread across the range of distribution**

URL: stats interview questions