Statistics Q&A

1. ***What is the difference between descriptive and inferential statistics?***

Descriptive Statistics is the method of organizing, summarizing, and presenting a specific set of data. It uses tools like the **average (mean)** and the **middle value (median)** to describe the features of the data you have. Its purpose is only to report and describe the characteristics of the observed group.

Inferential Statistics is the method of using data from a small group (**sample**) to make **generalizations and predictions** about a larger group (**population**). It uses techniques like **Hypothesis Testing** to figure out how likely the conclusions are to be correct. Its purpose is to go beyond the available data to draw broader conclusions.

2. ***What are mean, median, and mode? When do you use each?***

Mean (The Average)

The **Mean** is the **arithmetic average** of a dataset. It is calculated by summing all the values and dividing the total by the number of values.

- **When to Use It:** The mean is the most common measure and should be used when the data is **symmetrical** and does not contain significant **outliers** (extreme values). It uses all data points in the calculation.

Median (The Middle Value)

The **Median** is the **middle value** in a dataset when the values are arranged in ascending or descending order. If there is an even number of data points, the median is the average of the two middle values.

- **When to Use It:** The median is preferred when the data is **skewed** (not symmetrical) or contains **outliers**. Since the median is only dependent on the position of the values, it is not affected by extreme high or low numbers. For example, the median is typically used to report average **house prices** or **income**, as these datasets are often skewed by a few very high values.

Mode (The Most Frequent Value)

The **Mode** is the value that **appears most frequently** in a dataset. A dataset can have one mode, more than one mode (bimodal or multimodal), or no mode at all.

- **When to Use It:** The mode is primarily used for **categorical data** (like types of cars, favorite colors, or city of residence) where you can't calculate an average. It tells you the most common category or outcome.

3. ***What is variance? What is standard deviation?***

Variance

The statistical measure that tells us the **average of the squared distances** between every number in a data group and the group's average (mean).

Standard Deviation

The Standard Deviation is the number calculated by taking the **square root of the Variance**. It measures the **typical distance** or **amount of deviation** that values in the data set fall from the mean. It is the most commonly used measure of spread because it is expressed in the **original units** of the data, making it easy to understand.

4. ***What is a normal distribution, and why is it important?***

A normal distribution is a type of continuous probability distribution that is symmetric and bell-shaped, where most of the data points cluster around the mean, and fewer are found as you move further away. In a normal distribution, the mean, median, and mode are all equal and located at the center.

It is important because: It forms the basis for many statistical methods, especially through the **Central Limit Theorem**, which allows us to apply inferential statistics even when the underlying data is not perfectly normal.

5. ***What is skewness? Explain left-skewed vs right-skewed distributions.***

Skewness measures the asymmetry (lopsidedness) of a data distribution. It indicates where the bulk of the values lie and where the extreme values (the tail) are pulling the distribution.

1. Right-Skewed (or Positive Skew)

In a right-skewed distribution, the long tail extends to the right side of the graph.

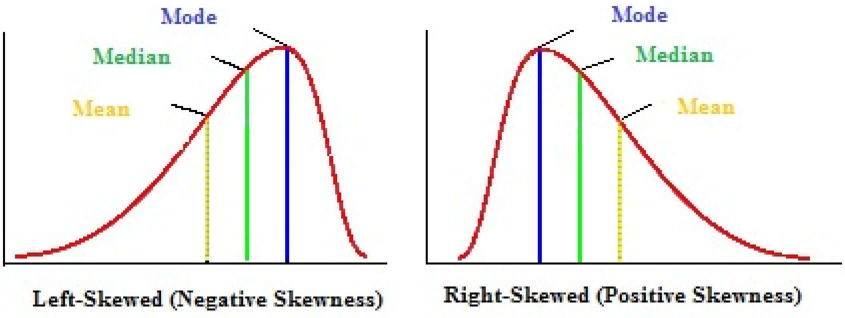
- The majority of the data points (the peak) are clustered on the left.
- The long tail on the right is caused by a few extremely large values dragging the Mean in that direction.
- Order of Measures: Mode<Median<Mean

2. Left-Skewed (or Negative Skew)

In a left-skewed distribution, the long tail extends to the left side of the graph.

- The majority of the data points (the peak) are clustered on the right.

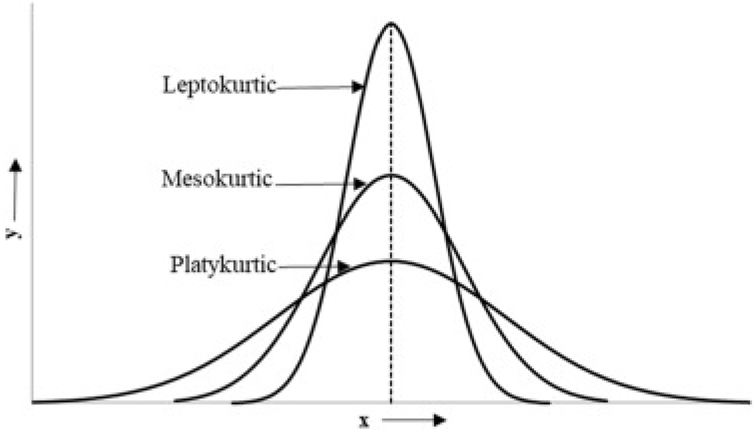
- The long tail on the left is caused by a few extremely small values dragging the Mean in that direction.
- Order of Measures: Mean<Median<Mode


Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

6. ***What is kurtosis, and what does it tell you?***

Kurtosis measures the **tailedness** of a distribution, telling you how much data is in the center peak and extreme tails compared to a normal curve.
It describes the curve's shape:

- **Leptokurtic** is tall with heavy tails (more outliers),
- **Platykurtic** is flat with thin tails (fewer outliers), and
- **Mesokurtic** is the standard bell curve.



7. ***what is null hypothesis?***

The Null Hypothesis (H0) is the foundational statement in statistical testing that assumes no effect, no difference, or no relationship exists between the variables being measured. It represents the status quo or the default assumption. The purpose of a statistical test is to gather enough evidence (a low p-value) to either reject this H0 in favor of the alternative hypothesis, or fail to reject it.

8. ***What is a p-value, and how do you interpret it?***

The p-value (probability value) is the key measure used in hypothesis testing to determine the statistical significance of a result. It represents the probability of observing your current data (or something more extreme) if the null hypothesis (the statement of no effect) were entirely true. We interpret the p-value by comparing it to a pre-determined threshold, α (usually 0.05). If p<α, we reject the null hypothesis, concluding the result is statistically significant and likely not due to random chance.

9. What is correlation vs causation?

Correlation means there is a relationship or association between two variables, and they tend to change together (e.g., as ice cream sales increase, so do sunburn cases). Causation means that a change in one variable directly produces a change in the other (e.g., turning on a light switch causes the light to turn on). The key takeaway is: Correlation does not imply causation; an observed relationship does not prove one thing caused the other.

10. What is the difference between population and sample?

The Population is the entire group you are interested in drawing conclusions about (e.g., all cars in a city). The Sample is a smaller, manageable, and representative subset of that population (e.g., 50 randomly selected cars from that city). We study the sample to make educated guesses (inferences) about the characteristics of the larger population.

11. What is the Central Limit Theorem (CLT)?

The Central Limit Theorem (CLT) is the most important theorem in statistics. It states that, regardless of the original population's shape, the distribution of the means of many large random samples will be approximately normally distributed (a Bell Curve). This is crucial because it allows statisticians to use the powerful properties of the Normal Distribution to make reliable predictions and inferences about a population from samples.

12. What is a Confidence Interval?

The Confidence Interval is a range of values that is likely to contain the true value of a population parameter (like the true population mean) based on data gathered from a sample. It provides a degree of certainty about the estimate, showing that if you repeated the sampling process many times, a certain percentage (e.g., 95%) of the intervals constructed would contain the true parameter.

13. What is a true parameter?

The True Parameter is the actual numerical characteristic or summary measure of the entire population that a researcher is interested in (e.g., the true average height of all adults in a country). It's a fixed value, but it is almost always unknown because it's impossible to measure every single member of the population. Statistical methods use sample data to estimate this true, unknown value.

14. What is a T-Test, and when is it typically used?

A T-Test is a type of inferential statistical test used to determine if there is a significant difference between the means of two groups. It's typically used when you have a relatively small sample size (usually less than 30) or, more importantly, when the population standard deviation is unknown. It allows researchers to determine if two groups are genuinely different or if any observed difference is simply due to random chance.

15. What is ANOVA (Analysis of Variance), and what is its purpose?

ANOVA (Analysis of Variance) is an inferential statistical test used to determine if there is a significant difference among the means of three or more independent groups simultaneously. Its core purpose is to compare the variability between the groups to the variability within the groups. By analyzing variances, ANOVA efficiently tells you if at least one group mean is statistically different from the others, which is much better than running many individual t-tests.

16. What is a Z-Score?

A Z-Score (or standard score) measures how many standard deviations (σ) a particular data point is away from the mean (μ) of its distribution. Its purpose is standardization: it allows you to compare values from completely different datasets (like an SAT score and an IQ score) by converting them to a common scale where the mean is 0 and the standard deviation is 1.

17. In simple terms, what are Degrees of Freedom (df)?

Degrees of Freedom (df) represent the number of independent pieces of information that went into calculating a statistic. A common formula is df=n−1, where n is the sample size.

18. What is a shape?

The shape of a distribution is a measure in statistics that describes the overall visual pattern or form of a dataset when it's graphed (usually as a histogram or density curve).

It is a key summary measure, alongside center (mean, median) and spread (standard deviation), used to understand the underlying characteristics of the data.

Key Features of Shape:

1. Symmetry: Is the distribution a mirror image on both sides of the center (like the Normal Distribution)?
2. Skewness: Is the data asymmetrical, with a long "tail" stretching to one side (left-skewed or right-skewed)?
3. Kurtosis: How peaked is the distribution, and how thick are its tails (are there more or fewer extreme outliers than expected)?
4. Modality: How many major peaks or "mounds" does the distribution have (unimodal, bimodal, etc.)?

Understanding the shape is critical because it dictates which statistical tests (parametric or non-parametric) are appropriate for further analysis.

19. What is a Type I Error?

A Type I Error occurs when a researcher incorrectly rejects the null hypothesis (H0) when the null hypothesis is actually true. It is often referred to as a false positive. The probability of committing a Type I Error is determined by the alpha (α) level (e.g., 0.05). This means you conclude that a significant effect exists (like a drug working) when, in reality, there is no effect.

20. What is a Type II Error?

A Type II Error occurs when a researcher fails to reject the null hypothesis (H0) when the null hypothesis is actually false. It's often referred to as a false negative. The probability of committing a Type II Error is represented by Beta (β). This means you conclude that no significant effect exists (e.g., the drug isn't working) when, in reality, a real effect is actually present.

21. What is Statistical Power?

Statistical Power is the probability that a statistical test will correctly reject the null hypothesis (H0) when the null hypothesis is actually false.

In simpler terms, power is a test's ability to detect a real effect or a real difference when one truly exists. It is calculated as $1-\beta$, where β is the probability of committing a Type II Error (a false negative). A high statistical power (often set at 0.80 or 80%) is desirable, as it means the test is unlikely to miss a true result.

22. What are Parametric and non-parametric Tests?

Parametric tests are statistical tests that make specific assumptions about the parameters of the population distribution, primarily that the data is drawn from a population that is normally distributed and has a known variance. These tests (like the t-test and ANOVA) are generally more powerful.
Non-parametric tests (like the Mann-Whitney U test) are used when these assumptions are violated or when dealing with ordinal or nominal data, as they do not rely on the population having a specific shape.

23. What is Simple Linear Regression?

Simple Linear Regression is a statistical method used to model the relationship between two continuous variables: one **dependent variable** (Y) and one **independent variable** (X).

Its primary purpose is to find the straight line that best fits the data, allowing you to **predict** the value of the dependent variable based on the value of the independent variable. The results are summarized by an equation: $Y=a+bX$, where 'a' is the y-intercept and 'b' is the slope (the regression coefficient).

24. What are variables?

A variable is simply a measurable characteristic, attribute, or quantity that can take on different values. In statistics and research, variables are the concepts being studied, analyzed, or modeled.

Variables are essential because they represent the data you collect, and they fall into two primary categories in a regression context:

- Independent Variable (X): This is the variable that is thought to be the cause or the input. You use its value to predict the outcome. (Also called the predictor).
- Dependent Variable (Y): This is the variable that is the effect or the output. It is the value you are trying to predict or explain. (Also called the response variable).

25. What is covariance ?

Covariance is a statistical measure that describes how two variables change together (or co-vary). It indicates the direction of the linear relationship between two variables, but not the strength.

- Positive Covariance: The two variables tend to increase or decrease together.
- Negative Covariance: As one variable increases, the other tends to decrease.
- Zero/Near-Zero Covariance: The two variables are independent or have no linear relationship.

Common, mandatory Interview questions for data analyst.

What is the difference between mean, median, and mode? When would you use each?

What is variance and standard deviation? Why are they important in data analysis?

What is a normal distribution, and why is it important in statistics?

What is skewness? Explain left-skewed and right-skewed distributions with examples.

What is correlation, and how is it different from causation?

What is a p-value, and how do you interpret it?

What is hypothesis testing, and why is it used?

What are confidence intervals, and how do you interpret them?

What are outliers, and how would you detect and handle them in data analysis?

What is sampling, and what are different types of sampling methods? Why is sampling important?

What is the Central Limit Theorem, and why is it important for data analysis?