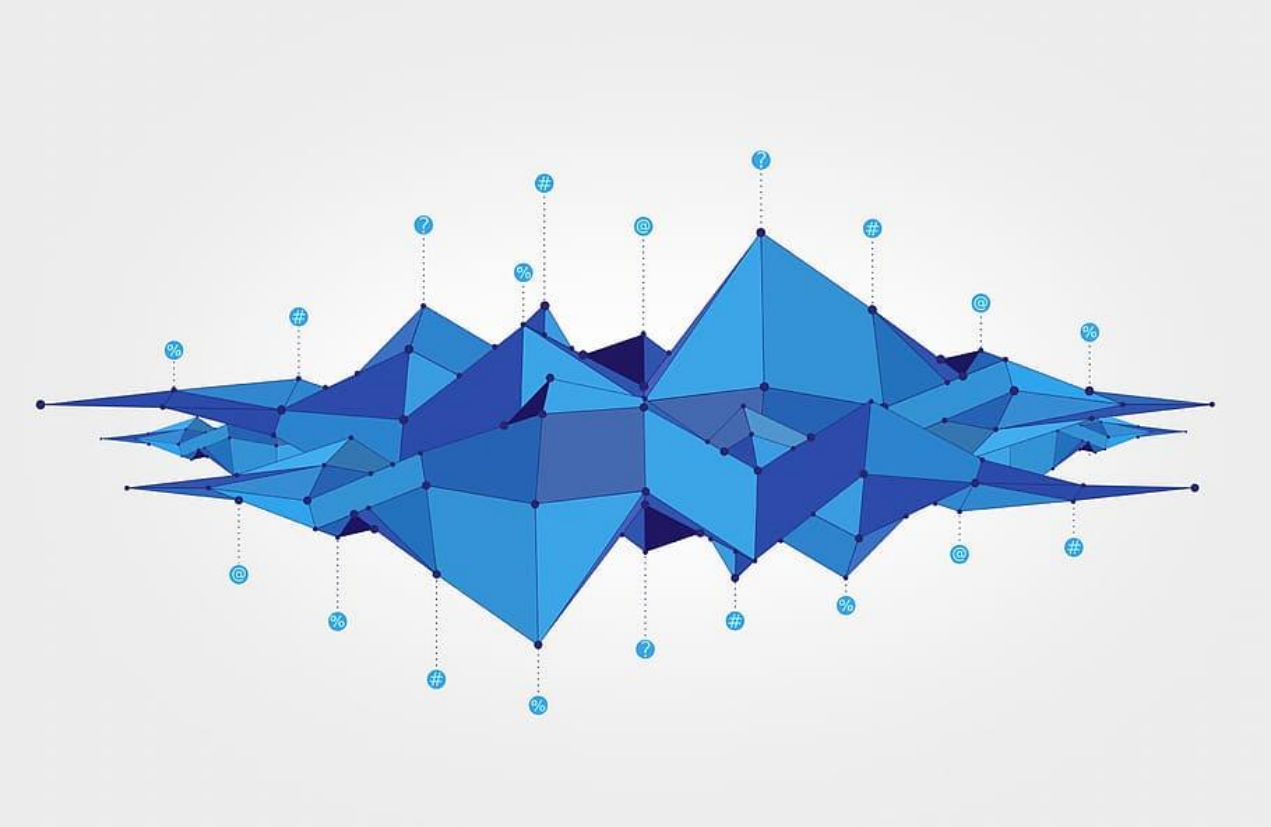


Top-100 Statistical Questions – By Pranjal Sharma ~(LinkedIn)

Unlock Your Success in ML Interviews with Our Comprehensive Guide!

Are you ready to ace those challenging machine learning interviews? Look no further! Welcome to the ultimate resource for mastering the statistical aspects of ML interviews.

🚀 Interview Oriented Solutions: Our meticulously crafted guide is tailor-made for your ML interview success. We've distilled the vast realm of statistics into a concise, interview-focused format, so you can confidently tackle any statistical question that comes your way.

📚 Short but Precise: Time is precious, and we respect that. Our guide gets straight to the point, providing you with the essential knowledge you need, without unnecessary fluff. Say goodbye to long-winded explanations and hello to efficient learning.

🎥 Video Solutions: We've taken learning to the next level by providing YouTube 🔗 References for video solutions. Visual learners rejoice! Now you can watch and learn alongside our detailed explanations.

🔥 Bonus Tips: What's an interview without a few secret weapons? We've got you covered with 🔥 Bonus Tips that will give you the edge over

other candidates. Discover insider strategies that can make all the difference.

📊 Examples & Formulas: Practicality is key. We present complex statistical concepts with real-world examples and essential formulas, ensuring you not only understand the theory but can apply it in interviews.

📋 Tabular Pointwise Differences: Say goodbye to confusion! We've organized key differences in tabular format, making it easy for you to quickly grasp the nuances of various statistical concepts.

Don't leave your success to chance. Prepare comprehensively and strategically with our "Top 100 Statistics Questions for ML Interviews" guide. Your dream job in machine learning awaits! 💼🌟

---

1. What are the most important topics in statistics?

Ans- I would say the most important topics in statistics for me are:

- Probability: This is the foundation of statistics, and it is essential for understanding many ML concepts, such as Bayesian inference and likelihood functions.
- Descriptive statistics: This is the process of summarizing and describing data. It includes basic measures such as mean, median, and mode, as well as more advanced measures such as standard deviation and correlation.
- Hypothesis testing: This is the process of using data to test a hypothesis about a population. It is used in ML to evaluate the performance of ML models and to select the best model for a given problem.
- Linear regression: This is a statistical model that can be used to predict a continuous variable from one or more other variables. It is a common ML algorithm that is used in a wide variety of applications.
- Logistic regression: This is a statistical model that can be used to predict a binary variable (e.g., yes/no, 1/0) from one or more other variables. It is another common ML algorithm that is used in many applications.

🔗 Ref- https://youtu.be/7V5jtI-ihm0?si=u2zB1HvQtRiHDBVz

2. What is exploratory data analysis?

Ans- Exploratory data analysis (EDA) is the process of analyzing and investigating data sets to identify patterns, trends, and anomalies. It is an essential step in any machine learning project, as it helps data scientists to better understand their data and to choose the right ML algorithms and parameters.

EDA can be performed using a variety of methods, including:

- Visualizing the data: This can be done using charts and graphs to get a better understanding of the distribution of the data and to identify any potential outliers.
- Calculating summary statistics: This includes calculating measures such as the mean, median, mode, standard deviation, and correlation.
- Performing statistical tests: This can be done to test hypotheses about the data and to identify any significant relationships between variables.

EDA is an iterative process, and it is important to go back and forth between different methods until you have a good understanding of your data.

🔗 **Ref-** https://m.youtube.com/watch?v=_Wp_b8wZ8Rc

3. What are quantitative data and qualitative data?
Ans-

- Quantitative data consists of numerical values that can be measured and expressed in terms of numbers. It deals with quantities and can be subjected to mathematical operations. Examples include age, height, weight, and temperature.

- Qualitative data, on the other hand, is categorical or descriptive in nature and represents qualities or characteristics that cannot be measured numerically. It is often 🔗 Referred to as categorical data. Examples include colors, types of animals, or customer feedback categories (e.g., "good," "bad," "neutral").
  🔗 **Ref** **Qualitative Vs Quantitative**

4. What is the meaning of KPI in statistics?
Ans- KPI stands for "Key Performance Indicator" in statistics. It's a measurable metric that is used to evaluate and gauge the performance or effectiveness of a particular process, system, or activity. KPIs are essential in various fields, including business, healthcare, and data analysis, to track progress and make informed decisions.
Steps to determine KPI-

1. Define the Business Goals
2. Define our key visual
3. Determine the measurements
4. Finalize your KPIs

🔗 Ref- KPI

5. What Is the Difference Between Univariate, Bivariate, and Multivariate Analysis?

Ans- Univariate, bivariate, and multivariate analysis are three different types of statistical analysis that differ in the number of variables they consider.

- **Univariate analysis** examines a single variable at a time. It is used to describe the distribution of a variable, such as its central tendency, variability, and shape. Common univariate statistics include mean, median, mode, standard deviation, and range.
- **Bivariate analysis** examines the relationship between two variables. It is used to determine if there is a correlation between the two variables and, if so, the strength and direction of the correlation. Common bivariate statistics include Pearson's correlation coefficient and Spearman's rank correlation coefficient.

- **Multivariate analysis** examines the relationships between three or more variables. It is used to identify complex patterns in data and to understand how multiple variables interact with each other. Common multivariate statistical methods include regression analysis, principal component analysis, and factor analysis.

Here is a short but lit and to-the-point explanation of the difference between univariate, bivariate, and multivariate analysis:

- Univariate analysis: One variable, one story.
- Bivariate analysis: Two variables, two stories.
- Multivariate analysis: Three or more variables, endless stories.

🔗 **Ref- [univariate, bivariate, and multivariate analysis]**

6. How Would You Approach a Dataset That's Missing More Than 30 Percent of Its Values?

Ans- There are two main approaches to handling missing values in a dataset: imputation and deletion.

Imputation is the process of filling in missing values with estimated values. There are many different imputation methods, each with its own advantages and disadvantages. Some common imputation methods include:

- Mean imputation: Replaces missing values with the mean value of the variable.
- Median imputation: Replaces missing values with the median value of the variable.
- Mode imputation: Replaces missing values with the most common value of the variable.
- K-nearest neighbors imputation: Replaces missing values with the average value of the k most similar observations.
- Regression imputation: Uses a regression model to predict missing values based on the other variables in the dataset.

Deletion is the process of removing rows or columns from the dataset that contain missing values. This is a simpler approach than imputation, but it can lead to a loss of information.

The best approach to handling missing values depends on the specific dataset and the intended use of the data. For a dataset with more than 30 percent missing values, I would typically use a combination of imputation and deletion.

First, I would try to understand the pattern of missing values. Are there any columns or rows that have a high percentage of missing values? Are there any groups of variables that are more likely to have missing values?

Once I have a better understanding of the missing values, I can decide which imputation methods to use. For example, if I have a column with a lot of missing values, I might use a k-nearest neighbors imputation method to fill in the missing values. If I have a row with a lot of missing values, I might delete the entire row.

After imputing and deleting missing values, I would evaluate the quality of the dataset. I would look at the distribution of the variables and check for any outliers. I would also train a machine learning model on the dataset and evaluate its performance.

🔗 **Ref- [Handling Missing Value]**

7. Give an example where the median is a better measure than the mean

Ans- One example is when the data is skewed, meaning that there are a few extreme values that are much higher or lower than the rest of the data. In this case, the mean can be pulled in the direction of the extreme values, giving a misleading impression of the central tendency of the data.

For example, consider the following data set:

```
[1, 2, 3, 4, 5, 100]
```

The mean of this data set is 18.33. However, this is not a good representation of the central tendency of the data, because the median is 3.5. This is because the extreme value of 100 is pulling the mean up.

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

Ans. Descriptive statistics describe some sample or population.

Inferential statistics attempts to infer from some sample to the larger population.

9. What are descriptive statistics?

Distribution – 🔗 Refers to the frequencies of responses.
Central Tendency – gives a measure or the average of each response.
Variability – shows the dispersion of a data set.

10. Can you state the method of dispersion of the data in statistics?

Ans- The most common methods of dispersion of data in statistics are:

- Range: The range is the difference between the largest and smallest values in a data set. It is a simple and easy-to-understand measure of dispersion, but it can be misleading if the data set has outliers.
- Variance: The variance is a measure of how spread out the data values are from the mean. It is calculated by taking the average of the squared deviations from the mean. The variance is a more robust measure of dispersion than the range, but it can be difficult to interpret on its own.
- Standard deviation: The standard deviation is the square root of the variance. It is a measure of how spread out the data values are from the mean, expressed in the same units as the data. The standard deviation is a very useful measure of dispersion, as it is easy to interpret and can be used to compare data sets from different populations.

🔗 Ref- Measures of Dispersion

11. How can we calculate the range of the data?

Ans- To calculate the range of a dataset, follow these steps:

Find the Maximum (Max) Value: Identify the highest value (largest number) in your dataset.

Find the Minimum (Min) Value: Identify the lowest value (smallest number) in your dataset.

Calculate the Range: Subtract the minimum value (Min) from the maximum value (Max).

Mathematically, the range (R) is calculated as:

R = Max - Min

The range provides a simple measure of the spread or variability in your data. It represents the difference between the highest and lowest values in the dataset.

$$\text{Range}(X) = \text{Max}(X) - \text{Min}(X)$$

12. Is the range sensitive to outliers?

Ans- Yes, the range is sensitive to outliers. The range is calculated as the difference between the highest and lowest values in a dataset. If there is an outlier in the dataset, it will either be the highest or lowest value, which will cause the range to be larger than it would be without the outlier.

Here is a simple example to illustrate this:

```
Dataset: 1, 2, 3, 4, 5, 100

Range: 100 - 1 = 99
```

Now, let's add an outlier to the dataset:

```
Dataset: 1, 2, 3, 4, 5, 100, 1000

Range: 1000 - 1 = 999
```

As you can see, the range has increased significantly from 99 to 999, simply because of the addition of a single outlier.

13. What is the meaning of standard deviation?

Ans. Standard deviation is a statistic that measures the dispersion of a dataset relative to its mean. It is the average amount of variability in your dataset. It tells you, on average, how far each value lies from the mean.
A high standard deviation means that values are generally far from the mean, while a low standard deviation indicates that values are clustered close to the mean.
The standard deviation is calculated as the square root of variance by determining each data point’s deviation relative to the mean.

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

14. What are the scenarios where outliers are kept in the data?

Ans- Outliers are data points that are significantly different from the rest of the data. They can be caused by errors in data collection or measurement, or they may represent genuine but rare events.

There are several scenarios where it is important to keep outliers in the data:

- When outliers may represent important information. For example, in a dataset of customer spending habits, an outlier may represent a customer who has made a very large purchase. This information could be valuable to the business, as it may indicate a new market opportunity or a change in customer behavior.
- When removing outliers may skew the results of the analysis. For example, if you are trying to identify the average salary of software engineers, removing outliers could result in an underestimate of the average salary. This is because outliers tend to be higher values, and removing them would lower the overall average.
- When the outliers are not caused by errors. For example, if you are tracking the number of visitors to a website, there may be days on which the number of visitors is significantly higher than normal. This could be due to a special promotion or a news story about the website. In this case, it would be important to keep the outliers in the data in order to get an accurate picture of the website's traffic over time.

15. What is Bessel's correction?

In statistics, Bessel's correction is the use of n-1 instead of n in several formulas, including the sample variance and standard deviation, where n is the number of observations in a sample. This method corrects the bias in the estimation of the population variance. It also partially corrects the bias in the estimation of the population standard deviation, thereby, providing more accurate results.

🔗 Ref- Bessel's correction

16. What do you understand about a spread out and concentrated curve?

Ans- A spread out curve in machine learning is one where the data points are widely distributed over a range of values. This means that there is a lot of variation in the data, and it is difficult to predict individual values with accuracy. A concentrated curve, on the other hand, is one where the data points are clustered closely together around a central

value. This means that there is less variation in the data, and it is easier to predict individual values with accuracy.

Here are some examples of spread out and concentrated curves:

- Spread out curve: The heights of students in a classroom.
- Concentrated curve: The weights of coins.

Spread out curves are more difficult to model than concentrated curves, because there is more variation in the data. However, spread out curves can also be more informative, because they provide more information about the range of values that the data can take.

Here are some of the implications of spread out and concentrated curves in machine learning:

- Spread out curves:
  - More difficult to train machine learning models.
  - Models may be less accurate at predicting individual values.
  - Models may be more robust to noise in the data.
- Concentrated curves:
  - Easier to train machine learning models.
  - Models may be more accurate at predicting individual values.
  - Models may be more sensitive to noise in the data.

17. Can you calculate the coefficient of variation?

Ans- The coefficient of variation (CV) is a measure of relative variability. It is calculated by dividing the standard deviation by the mean and multiplying by 100 to express the result as a percentage.

The formula for the CV is:

```
CV = (Standard deviation / Mean) * 100
```

The CV is a useful statistic for comparing the variability of different datasets, even if they have different units of measurement.

	Coefficient of Variation	Standard Deviation
Population	$\frac{\sigma}{\mu} \times 100$	$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$
Sample	$\frac{s}{\mu} \times 100$	$s = \sqrt{\frac{\sum(x_i - \mu)^2}{N - 1}}$

For example, you could use the CV to compare the variability of the heights of two different groups of people, even if one group of people is measured in feet and the other group of people is measured in centimeters.

🔗 Ref- Cofficent of Variation

18. State the case where the median is a better measure when compared to the mean.

Ans- The median is a better measure of central tendency when the data is skewed or contains outliers. Skewed data means that the data is not evenly distributed, with more values on one side of the distribution than the other. Outliers are extreme values that are far from the rest of the data.

The median is less affected by skewed data and outliers than the mean. For example, consider the following data set:

```
[1, 2, 3, 4, 5, 6, 7, 8, 9, 1000]
```

The mean of this data set is 106, but the median is 5. The mean is skewed by the outlier, 1000. The median, on the other hand, is more representative of the typical value in the data set.

For example, consider a dataset of salaries within a small company. Most employees have salaries within a typical range, but there is one executive with an exceptionally high salary. In this case, the mean salary would be heavily influenced by the outlier, giving a distorted view of the typical salary within the company. The median, however, would provide a more accurate representation of the central salary value because it is not affected by extreme values.

🔗 Ref- Median is better that mean

19. How is missing data handled in statistics?

Ans- Missing data is a common problem in statistical analysis. It can occur for a variety of reasons, such as participants withdrawing from a study, failing to answer certain questions, or data entry errors.

There are a number of different ways to handle missing data. The best approach depends on the specific situation and the type of analysis being performed. Some common methods include:

- Listwise deletion: This involves removing all cases from the analysis that have any missing data. This is the simplest method, but it can lead to a loss of power and bias in the results, especially if the amount of missing data is high.
- Mean imputation: This involves replacing missing values with the mean value of the variable for all other cases. This is a relatively simple method, but it can underestimate the variability of the data and lead to biased results if the missing values are not randomly distributed.
- Multiple imputation: This involves creating multiple copies of the dataset and imputing the missing values in each copy using a different imputation model. The results of the analysis are then combined to produce a more accurate and unbiased estimate. This is a more complex method, but it is generally considered to be the best way to handle missing data.

🔗 **Ref- Handle the missing Data**

20. What is meant by mean imputation for missing data? Why is it bad?

Ans- Mean imputation is a method of filling in missing values in a dataset by replacing them with the mean value of the column in which they are missing. It is a simple and straightforward method, but it has a number of drawbacks.

Why is mean imputation bad?

- It can bias the results of any analysis performed on the data, especially if the missing data is not missing completely at random (MCAR).
- It can underestimate the variability of the data, as it replaces missing values with a single value.
- It can obscure patterns in the data, as it replaces missing values with a value that is not necessarily representative of the true value.

  Example

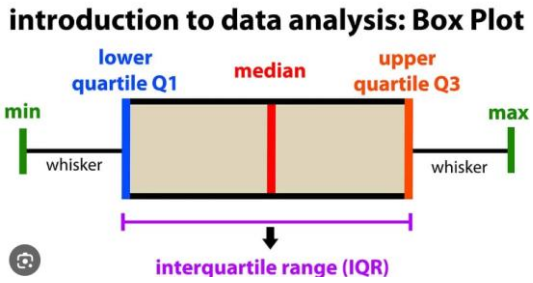
- Suppose we have a dataset of student heights, with some missing values. We could use mean imputation to fill in the missing values by replacing them with the mean height of all the students in the dataset. However, this would bias the results of any analysis performed on the data, as it would overestimate the height of the students with missing values.

- A better approach would be to use a more sophisticated imputation method, such as multiple imputation, which takes into account the uncertainty around the missing values.

🔗 **Ref- [Two Big Problems with Mean Imputation]**

21. What is the benefit of using box plots?

Ans- Box plots are a great way to visualize data distribution and identify outliers. They are easy to interpret, concise, informative, versatile, Comparison of Multiple Distributions, Resistant to Extreme Values, Quick Assessment of Skewness and Space-Efficient.
They are especially useful for comparing data from different groups or over time.



Here are some of the benefits of using box plots:

- Easy to interpret: Box plots are very visually appealing and easy to interpret, even for non-statisticians.
- Concise: Box plots can convey a lot of information about a data set in a very concise way.
- Informative: Box plots can show the central tendency, spread, and skewness of a data set. They can also identify outliers.
- Versatile: Box plots can be used to compare data from different groups or over time.

🔥 Bonus tip: If you are interviewing for a job in machine learning, be sure to mention that box plots can be used to identify outliers in training data. This is important because outliers can skew the results of machine learning models.
🔗 Ref- Box Plot

22. What is the meaning of the five-number summary in Statistics?

Ans- The five-number summary is a statistical method that provides a concise overview of a data set by summarizing its central tendency, spread, and skewness. It is calculated by ordering the data set from smallest to largest and then identifying the following five values:

- Minimum: The smallest value in the data set.
- First quartile (Q1): The median of the lower half of the data set.
- Median (Q2): The middle value in the data set.
- Third quartile (Q3): The median of the upper half of the data set.
- Maximum: The largest value in the data set.

The five-number summary can be used to create a box plot, which is a visual representation of the data distribution. The box plot shows the median, quartiles, and outliers, and it can be used to quickly identify important features of the data set, such as its central tendency, spread, and skewness.

🔥 Bonus tip: In short, the five-number summary is a powerful statistical tool that can be used to quickly and easily understand the distribution of a data set.

🔗 Ref- Box Plot- 5 number summary

23. What is the difference between the First quartile, the IInd quartile, and the IIIrd quartile?

Ans- The first quartile (Q1), second quartile (Q2), and third quartile (Q3) are all measures of central tendency, but they provide different levels of detail.

First quartile: The first quartile is the middle value of the lower half of the data set. It represents the value below which 25% of the data falls.

Second quartile: The second quartile is the median value of the data set. It represents the value below which 50% of the data falls.

Third quartile: The third quartile is the middle value of the upper half of the data set. It represents the value below which 75% of the data falls.

The Quartile Formula for Q1 = $\frac{1}{4}(n+1)^{th}$ term

The Quartile Formula for Q3 = $\frac{3}{4}(n+1)^{th}$ term

The Quartile Formula for Q2 = Q3 – Q1 (Equivalent to Median)

24. What is the difference between percent and percentile?

Ans- Percent and percentile are related concepts but serve different purposes in statistics.

Percent:

A percent is a unit of measurement that represents a fraction of 100. It is often denoted by the symbol "%."
Percentages are used to express proportions, ratios, or relative sizes. For example, if you score 80 out of 100 on a test, you've achieved 80 percent.

Percentile:

A percentile is a statistical measure used to describe the position of a particular value within a dataset. Percentiles divide a dataset into 100 equal parts (hence the "centile" part of the word). When you say you're in the 90th percentile in a test score, it means your score is higher than or equal to 90% of the scores in the dataset.

Percentiles help understand where a specific data point stands relative to others in the distribution. For instance, if you're in the 90th percentile for income, it means your income is higher than or equal to 90% of the incomes in a population.

🔗 Ref- Percentage & percentile

25. What is an Outlier?

Ans- An outlier is a data point that is significantly different from the rest of the data set. It can be much higher or much lower than the other values. Outliers can be caused by errors in data collection or entry, but they can also be real values that represent unusual or unexpected events.

Example:

Consider the following data set of test scores:

```
90, 85, 88, 92, 75, 100
```

The score of 75 is an outlier because it is much lower than the other scores. It could be caused by a mistake, such as the student not feeling well on the day of the test, or it could be a real value that represents a student who is struggling academically.

Identifying outliers

There are a number of ways to identify outliers in a data set. One common method is to use the interquartile range (IQR). The IQR is the difference between the third and first quartiles of the data. Outliers are defined as any values that are more than 1.5 IQRs above or below the median.

Another method for identifying outliers is to use a boxplot. A boxplot is a graphical representation of the data that shows the median, IQR, and outliers. Outliers are typically plotted as individual points outside of the box.

🔗 Ref- What is an Outlier?

26. What is the impact of outliers in a dataset?

Ans- Outliers are data points that fall significantly outside the main pattern of the data. They can have a significant impact on statistical analysis, as they can skew the results and make it difficult to identify trends and patterns.

Impact of outliers:

- Mean and median: Outliers can have a large impact on the mean and median of a dataset, especially if the dataset is small. This is because the mean and median are calculated by averaging all the data points, including the outliers.
- Standard deviation: Outliers can also have a large impact on the standard deviation of a dataset. The standard deviation is a measure of how spread out the data is, and it is calculated by taking the square root of the variance. The variance is calculated by averaging the squared deviations of each data point from the mean. Outliers can increase the variance and the 🔗 Refore the standard deviation.
- Machine learning models: Outliers can also have a negative impact on machine learning models. Machine learning models are trained on data, and if the data contains outliers, the model will learn to fit the outliers instead of the main pattern of the data. This can lead to the model making inaccurate predictions.

🔥 Bonus tip: Outliers can have a significant impact on statistical analysis and machine learning models. It is important to be aware of the potential impact of outliers and to take steps to mitigate their impact. Steps to mitigate the impact of outliers include:

- Identifying outliers

- Removing outliers
- Using robust statistical methods
- Using machine learning algorithms that are robust to outliers

🔗 Ref- [What is an Outlier?]

27. Mention methods to screen for outliers in a dataset.

Ans- When screening for outliers in a dataset, you can use various methods. Here are some common approaches:

Visual Inspection:
Box plots: Box plots visually display the distribution of data, making it easy to spot outliers as points outside the "whiskers."
Scatter plots: Scatter plots can reveal outliers as data points that fall far away from the main cluster of points.

Summary Statistics:
Z-Score: Calculate the Z-Score for each data point, which measures how many standard deviations it is away from the mean. Points with high absolute Z-Scores are potential outliers.
IQR (Interquartile Range): Calculate the IQR, and data points falling below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR are considered outliers.

Machine Learning Models:
Some machine learning models, like Isolation Forests and One-Class SVMs, are designed to identify outliers in data.

Domain Knowledge:
Use domain knowledge to identify values that are logically or practically improbable, given the context of the data.
🔗 Ref- [Screening the outliers]

28. How you can handle outliers in the datasets.

Ans- There are a few ways to handle outliers in datasets, depending on the specific situation:

1. Remove the outliers. This is the simplest approach, but it should only be done if the outliers are clearly erroneous, such as a data entry error.

2. Winsorize the outliers. This involves setting the outliers to the next highest or lowest non-outlier value. This can be useful if the outliers are not necessarily erroneous, but they are still skewing the results.

3. Transform the data. This can involve taking the logarithm of the data, or using another transformation to make the distribution more normal. This can be useful if the outliers are caused by the data having a skewed distribution.

4. Use a robust statistical method. There are a number of statistical methods that are designed to be robust to outliers. These methods can be used to estimate the mean, median, and other statistical measures of a dataset without being unduly influenced by outliers.

🔥 Bonus tip: It is important to first understand why the outliers are occurring. Are they due to data entry errors, or are they legitimate data points? Once you understand the cause of the outliers, you can choose the appropriate method for handling them.

It is also important to consider the impact of handling outliers on your results. For example, if you remove too many outliers, you may introduce bias into your dataset. If you transform the data, you may need to be ca🔗 Reful to choose a transformation that does not introduce artifacts into the data.

Finally, it is important to document how you handled outliers in your analysis. This will help you and others to understand the results of your analysis, and to reproduce your work.

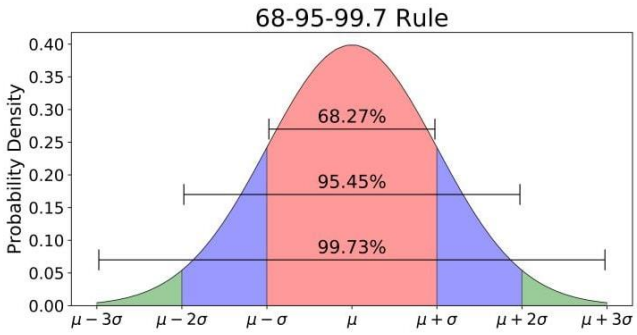
🔗 Ref- Handling of Outlier

29. What is the empirical rule?

Ans- The empirical rule, also known as the 68-95-99.7 rule, states that for a normally distributed dataset, approximately:

- 68% of the data falls within 1 standard deviation of the mean
- 95% of the data falls within 2 standard deviations of the mean
- 99.7% of the data falls within 3 standard deviations of the mean

This can be visualized as a bell-shaped curve, with the mean at the center and the standard deviations on either side. The greater the number of standard deviations away from the mean, the smaller the percentage of data that falls within that range.



🔗 Ref- Empirical rule

30. How to calculate range and interquartile range?

Ans- Range is the difference between the largest and smallest values in a dataset. It is calculated as follows:

```
Range = Maximum value - Minimum value
```

Interquartile range (IQR) is the difference between the third quartile (Q3) and the first quartile (Q1). It is a measure of the spread of the middle 50% of the data. It is calculated as follows:

```
IQR = Q3 - Q1
```

To calculate the quartiles, you can use the following steps:

1. Order the data in ascending order.
2. Find the median, which is the middle value in the ordered dataset.
3. The first quartile (Q1) is the median of the lower half of the dataset.
4. The third quartile (Q3) is the median of the upper half of the dataset.

Example:

```
Dataset: 1, 2, 3, 5, 7, 11, 13, 17, 19, 23, 29

Ordered dataset: 1, 2, 3, 5, 7, 11, 13, 17, 19, 23, 29
```

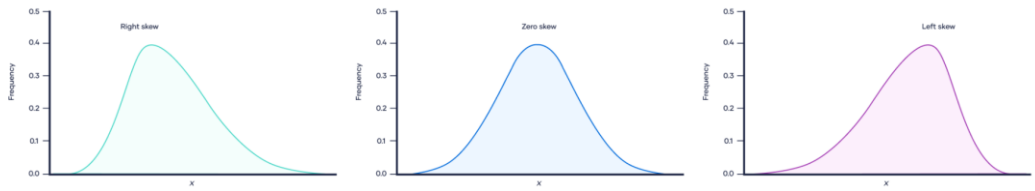
```
Median: (13 + 17) / 2 = 15

Q1: (3 + 5) / 2 = 4

Q3: (19 + 23) / 2 = 21

Range: 29 - 1 = 28

IQR: Q3 - Q1 = 21 - 4 = 17
```

31. What is skewness?

Ans- Skewness is a measure of the asymmetry of a probability distribution. It is a measure of how much the distribution is tilted to one side or the other. A perfectly symmetrical distribution has a skewness of 0. A positive skewness means that the distribution is tilted to the left, with a longer tail on the right. A negative skewness means that the distribution is tilted to the right, with a longer tail on the left.

Example:

Imagine a dataset of income. If the dataset is perfectly symmetrical, then the mean, median, and mode will all be equal. However, if the dataset is skewed, then the mean, median, and mode will be different. For example, if the dataset is positively skewed, then the mean will be greater than the median and the median will be greater than the mode. This is because the dataset has a longer tail on the right, meaning that there are more people with high incomes than low incomes.



32. What are the different measures of Skewness?

Ans- Skewness is a measure of the asymmetry of a probability distribution. It can be positive or negative, depending on whether the tail of the distribution is longer on the right or left side, respectively.

There are several different ways to measure skewness, but the most common are:

- Pearson's first coefficient of skewness: This measure is calculated by subtracting the mean from the median and then dividing by the standard deviation.

  ```
  Skewness = (Mean - Median) / Standard deviation
  ```

- Bowley's coefficient of skewness: This measure is calculated by subtracting the third quartile from the first quartile and then dividing by the interquartile range.

  ```
  Skewness = (Q3 - Q1) / IQR
  ```

- Fisher's coefficient of skewness: This measure is calculated by taking the natural logarithm of the ratio of the third quartile to the first quartile.

  ```
  Skewness = ln(Q3 / Q1)
  ```

All three of these measures of skewness range from -3 to 3, with a value of 0 indicating a perfectly symmetrical distribution. A positive value indicates a positively skewed distribution, while a negative value indicates a negatively skewed distribution.

🔗 Ref- [Measure of Skewness]

33. What is kurtosis?

Kurtosis is a measure of the peakedness of a probability distribution. It tells you whether the distribution is more peaked or flatter than a normal distribution.

Kurtosis is typically measured using Pearson's fourth coefficient of kurtosis, which is calculated by:

```
Kurtosis = (Fourth moment - 3(Standard deviation)^4) / (Standard deviation)^4
```

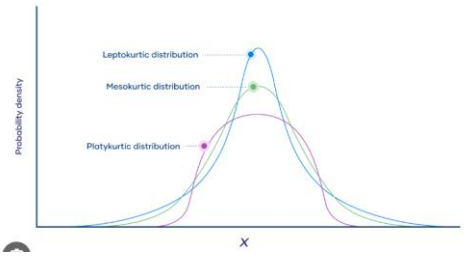
where:

- Fourth moment is the average of the fourth powers of the deviations from the mean
- Standard deviation is a measure of how spread out the data points are

Pearson's fourth coefficient of kurtosis is a unitless measure, and a value of 3 indicates a normal distribution. Values greater than 3 indicate a more peaked distribution, while values less than 3 indicate a flatter distribution.

There is 3 types of kurtosis

1. Leptokurtic
2. Mesokurtic
3. platykurtic



🔥 Bonus tip: Kurtosis is a useful measure for understanding the shape of a distribution and for identifying outliers. For example, a distribution with a high kurtosis may have more outliers than a distribution with a low kurtosis.

Kurtosis is also a useful measure in machine learning. For example, some machine learning algorithms, such as support vector machines, are sensitive to the shape of the distribution of the training data. By understanding the kurtosis of the training data, machine learning practitioners can choose the appropriate algorithm and tune its parameters accordingly.

🔗 Ref- [kurtosis]

34. Where are long-tailed distributions used?

Ans- Long-tailed distributions are used in various fields and applications where rare or extreme events play a crucial role. Here are some common areas where long-tailed distributions are used:

- Finance: Long-tailed distributions are employed in risk management and finance to model extreme events such as stock market crashes, large-scale financial losses, or extreme market volatility.
- Insurance: In the insurance industry, long-tailed distributions are used to assess and manage the risk associated with rare but high-impact events, such as natural disasters or catastrophic accidents.

- Internet and Social Media: Long-tailed distributions are observed in internet phenomena like website traffic, social media shares, and popularity of online content. They help in understanding the distribution of user engagement.
- Natural Disasters: Long-tailed distributions are used in modeling natural disasters like earthquakes, hurricanes, and floods, as they capture the infrequent but severe occurrences.
- Biomedical Research: Long-tailed distributions can be found in epidemiology when modeling the spread of diseases, as well as in genomics when analyzing the distribution of rare genetic mutations.
- Customer Behavior: In business analytics, long-tailed distributions are used to understand customer behavior, such as purchase frequency, where a small percentage of customers may contribute to a significant portion of revenue.

35. What is the central limit theorem?

Ans- The central limit theorem is a fundamental theorem in statistics that states that the distribution of the sample mean will approximate a normal distribution as the sample size increases, regardless of the distribution of the population from which the sample is drawn. This means that if we take a large enough sample of data from any population, the average of those samples will be approximately normally distributed. This is true even if the population itself is not normally distributed.

The central limit theorem is important because it allows us to make inferences about populations based on samples. For example, we can use the CLT to calculate confidence intervals and to conduct hypothesis tests.

Here is a simple example of how the central limit theorem can be used in machine learning:

Suppose we are training a machine learning model to predict the price of a house. The population of all house prices is likely to be non-normally distributed, with a long tail of high-priced houses. However, if we take a large enough sample of house prices, the distribution of those sample means will be approximately normally distributed. This means that we can use the CLT to calculate confidence intervals for our predictions. For example, we might say that we are 95% confident that the true price of the house is within $10,000 of our predicted price.

🔥 Bonus tip: The central limit theorem is a powerful tool that can be used to improve the accuracy and reliability of machine learning models.

🔗 Ref- Central Limit Theorem

36. Can you give an example to denote the working of the central limit theorem?

Ans- The Central Limit Theorem (CLT) is a fundamental concept in statistics that states that the distribution of the sample means of a large enough sample, drawn from any population, approaches a normal distribution, regardless of the population's underlying distribution.

Here's a simple example to illustrate the Central Limit Theorem:

Imagine you have a population of exam scores for a large group of students, and the scores in the population do not follow a normal distribution; they might be skewed or have any shape. Now, if you were to take multiple random samples of, let's say, 30 students' scores each, and calculate the mean score for each sample, you would notice that the distribution of these sample means starts to resemble a bell-shaped normal distribution as you take more and more samples.

This phenomenon occurs because of the Central Limit Theorem. It tells us that even if the original population's distribution is not normal, the distribution of sample means becomes approximately normal as the sample size increases. This normal distribution of sample means is incredibly useful in statistical inference and hypothesis testing.

Ref- Central Limit Theorem

37. What general conditions must be satisfied for the central limit theorem to hold?

Ans-

- The data must be sampled randomly
- The sample values must be independent of each other
- The sample size must be sufficiently large, generally it should be greater or equal than 30

38. What is the meaning of selection bias?

Ans- Selection bias is a type of bias that occurs when the sample of a study is not representative of the population that the study is trying to generalize to. This can happen for a variety of reasons, such as:

- Self-selection bias: This occurs when people volunteer themselves to participate in a study, which can lead to a sample that is over- or under-represented by certain groups.
- Convenience sampling: This occurs when researchers recruit participants from a convenient source, such as students in a class or patients in a hospital. This can lead to a sample that is not representative of the general population.
- Non-response bias: This occurs when some members of the sample do not respond to the study, which can lead to a sample that is not representative of the original population.

Selection bias can lead to misleading results, as the conclusions of the study may not be applicable to the population that it is trying to generalize to.

example of selection bias is a study that investigates the relationship between smoking and lung cancer. The study recruits participants from a group of people who are already being treated for lung cancer. However, only people who are still alive are eligible to participate. This leads to a sample that is under-represented by people who have died from lung cancer. As a result, the study may underestimate the risks of smoking.

To avoid selection bias-

There are a number of things that researchers can do to avoid selection bias, such as:

- Using probability sampling methods to recruit participants. This ensures that all members of the population have an equal chance of being selected for the study.
- Using response rates to assess the representativeness of the sample. If the response rate is low, it is possible that the sample is not representative of the population.
- Using statistical methods to adjust for selection bias. This can be done by weighting the data or using a statistical model to control for the effects of selection bias.

Ref- Selection Bias

39. What are the types of selection bias in statistics?

Ans- There are several types of selection bias, including:

- **Sampling Bias**: This occurs when the sample selected is not representative of the population because certain groups or individuals are more likely to be included or excluded. For example, conducting a phone survey during the day might exclude individuals who work during those hours, biasing the sample.
- **Non-Response Bias**: Non-response bias occurs when individuals selected for the sample do not respond, and their characteristics differ from those who do respond. This can lead to an unrepresentative sample.

- **Survivorship Bias**: Survivorship bias happens when the analysis is based only on data from individuals or items that have "survived" a particular process or selection criteria. It can lead to incorrect conclusions, especially in situations where the non-survivors have important information.

- **Healthy User Bias**: This occurs when a study includes only individuals who are healthy or have a particular characteristic, skewing the results. For example, if a health survey is conducted at a fitness center, it may not represent the health of the general population.

- **Publication Bias**: Publication bias arises when only significant or interesting results are published, while less exciting or negative results are not. This can create a skewed perception of the overall evidence on a particular topic.

🔗 **Ref- Type Of Biases**

40. What is the probability of throwing two fair dice when the sum is 8?

Ans- To calculate the probability of throwing two fair dice and getting a sum of 8, you need to consider the possible combinations that result in a sum of 8. There are five such combinations: (2, 6), (3, 5), (4, 4), (5, 3), and (6, 2).

Each die has six sides, so there are a total of 6 * 6 = 36 possible outcomes when rolling two dice.

The probability of getting a sum of 8 is the number of favourable outcomes (5) divided by the total number of possible outcomes (36).

Probability = (Number of Favourable Outcomes) / (Total Number of Possible Outcomes)
Probability = 5 / 36

🔗 Ref- Solution

41. What are the different types of Probability Distribution used in Data Science?

Ans- There are two main types of probability distributions: discrete and continuous.

- Discrete probability distributions are used to model events with a finite number of outcomes. For example, the probability of rolling a 6 on a die is a discrete probability distribution.
- Continuous probability distributions are used to model events with an infinite number of outcomes. For example, the height of a person is a continuous probability distribution.

Some of the most common probability distributions used in data science are:

- Normal distribution: The normal distribution is also known as the Gaussian distribution. It is a bell-shaped curve that is symmetrical around the mean. The normal distribution is used to model a wide variety of phenomena, such as human height, IQ scores, and test scores.
- Uniform distribution: The uniform distribution describes a situation where all outcomes are equally likely. For example, the probability of flipping a coin and getting heads or tails is a uniform distribution.
- Bernoulli distribution: The Bernoulli distribution describes a situation with two possible outcomes, such as success or failure. For example, the probability of a customer clicking on an ad is a Bernoulli distribution.
- Poisson distribution: The Poisson distribution describes the probability of a certain number of events occurring in a fixed interval of time or space. For example, the probability of receiving a certain number of emails in a day is a Poisson distribution.
- Exponential distribution: The exponential distribution describes the probability of an event occurring after a certain amount of time. For example, the probability of a customer returning a product after a certain number of days is an exponential distribution.

🔗 Ref- Probability Distribution

42. What do you understand by the term Normal Distribution or What is a bell-curve distribution?

Ans- The normal distribution, also known as the bell-curve distribution, is a type of probability distribution that is symmetrical and bell-shaped. It is the most common distribution in statistics and probability, and it is used to model a wide variety of real-world phenomena, such as human height, IQ scores, and test results.

The normal distribution has a number of important properties, including:

- It is symmetrical, meaning that the left and right halves of the curve are mirror images of each other.
- The mean, median, and mode of the distribution are all equal.
- 68% of the values in the distribution fall within one standard deviation of the mean.
- 95% of the values in the distribution fall within two standard deviations of the mean.

The normal distribution can be visualized as a bell-shaped curve, with the highest point of the curve at the mean. The curve then tails off to the left and right of the mean, with fewer and fewer values as you move further away from the mean.

🔗 Ref- Normal Distribution

43. Can you state the formula for normal distribution?

Ans- The normal distribution is a bell-shaped curve that describes how data is distributed. It is also known as the Gaussian distribution. The normal distribution is the most important distribution in statistics and machine learning.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

$f(x)$ = probability density function
σ = standard deviation
μ = mean

44. What type of data does not have a normal distribution or a Gaussian distribution?

Ans- Many types of data do not have a normal distribution, including:

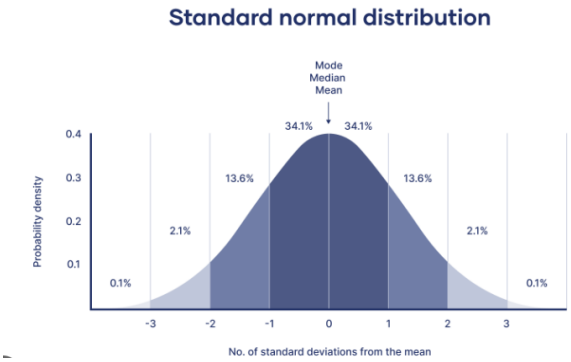
- **Count data**: This type of data represents the number of times something happens, such as the number of goals scored in a soccer game or the number of customers who visit a store on a given day. Count data typically follows a Poisson distribution or a binomial distribution.
- **Categorical data**: This type of data represents different categories, such as gender, product type, or customer satisfaction. Categorical data does not have a natural ordering and does not follow a normal distribution.
- **Skewed data**: This type of data has a long tail on one side of the distribution, with more values at one end than the other. Skewed data is common in income data, where there are a few very high earners and many people with lower incomes.
- **Truncated data**: This type of data has been cut off at some point, such as data on the number of children people have, where the maximum value is 20. Truncated data does not follow a normal distribution because it is missing some values from the tail of the distribution.

🔗 Ref-

45. What is the relationship between mean and median in a normal distribution?

Ans- **Mean and median are equal in a normal distribution.**

This is because a normal distribution is symmetrical around its mean, meaning that the left and right halves of the distribution are identical. This can be visualized by looking at the bell curve shape of a normal distribution:



🔥 Bonus tip: The mean and median are equal in a normal distribution because the distribution is symmetrical around its mean. This means that the left and right halves of the distribution are identical, so the middle value (the median) and the average value (the mean) will always be the same.

🔗 Ref- [mean and median in a normal distribution]

46. What are some of the properties of a normal distribution?

Ans- Some of the properties of a normal distribution:

- Symmetrical: The normal distribution is symmetrical around the mean, which means that the left and right halves of the curve are mirror images of each other.
- Bell-shaped: The normal distribution is bell-shaped, with the highest point of the curve at the mean and the tails of the curve tapering off towards infinity.
- Continuous: The normal distribution is a continuous distribution, which means that the random variable can take on any value within a certain range.
- 68-95-99.7 rule: Approximately 68% of the values in a normal distribution fall within 1 standard deviation of the mean, 95% fall within 2 standard deviations of the mean, and 99.7% fall within 3 standard deviations of the mean.

🔗 **Ref-Properties of Normal DIstribution**

47. What is the assumption of normality?

Ans- The assumption of normality is that the data is distributed according to a normal distribution, also known as a Gaussian distribution. This means that the data is bell-shaped, with the majority of values falling near the mean and fewer values falling further away.

The assumption of normality is important for many machine learning algorithms, such as linear regression and logistic regression. These algorithms use the normal distribution to calculate probabilities and make predictions. If the data is not normally distributed, the results of these algorithms may be inaccurate.

🔥 **Bonus tip:** key points about the assumption of normality in machine learning:

- Why is the assumption of normality important? The assumption of normality is important because many machine learning algorithms rely on the normal distribution to calculate probabilities and make predictions. If the data is not normally distributed, the results of these algorithms may be inaccurate.
- How to check if data is normally distributed? There are a few different ways to check if data is normally distributed. One common method is to create a histogram of the data. If the histogram is bell-shaped, then the data is likely to be normally distributed. Another method is to use a normality test, such as the Shapiro-Wilk test or the Anderson-Darling test.
- What to do if data is not normally distributed? If data is not normally distributed, there are a few different things you can do. One option is to transform the data so that it becomes more normally distributed. Another option is to use a machine learning algorithm that is robust to violations of the normality assumption, such as a decision tree or a random forest.

🔗 Ref- Assumption of normality

48. How to convert normal distribution to standard normal distribution?

Ans- To convert a normal distribution to a standard normal distribution, we use the following formula:

$$z = (x - \mu) / \sigma$$

where:

- z is the standard normal score
- x is the value in the original normal distribution
- μ is the mean of the original normal distribution
- σ is the standard deviation of the original normal distribution

The standard normal distribution has a mean of 0 and a standard deviation of 1. The 🔗 Refore, converting a normal distribution to a standard normal distribution allows us to compare different normal distributions on a common scale.

🔥 **Bonus tip:** Applications:

Converting a normal distribution to a standard normal distribution has many applications in statistics and machine learning. For example, it can be used to calculate the probability of a certain event occurring, or to compare the performance of different machine learning models.

🔗 Ref- Transformation to Normal distribution

49. Can you tell me the range of the values in standard normal distribution?

Ans- The range of values in the standard normal distribution is $-\infty$ to ∞. This means that any real number is possible, regardless of how large or small.

However, the vast majority of values in the standard normal distribution are close to the mean, which is 0. In fact, 68% of values fall within 1 standard deviation of the mean, 95% fall within 2 standard deviations of the mean, and 99.7% fall within 3 standard deviations of the mean.

🔥 **Bonus tip:** The standard normal distribution is a special type of normal distribution that has a mean of 0 and a standard deviation of 1. It is also known as the Z-distribution.

The standard normal distribution is very important in statistics because it allows us to compare different normal distributions, regardless of their mean and standard deviation. This is done by converting each normal distribution to a Z-distribution and then comparing the Z-scores.

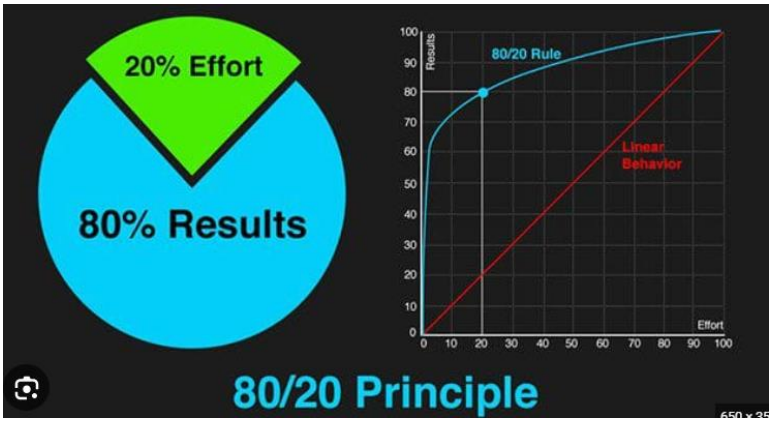
🔗 Ref- Range of Standard Normal distribution

50. What is the Pareto principle?

Ans- The Pareto principle, also known as the 80/20 rule, states that 80% of the effects come from 20% of the causes. It was first observed by Italian economist Vilfredo Pareto in the early 1900s, and has since been found to apply to a wide range of phenomena, including business, economics, and even nature.

For example, in business, the Pareto principle might be used to describe the fact that 20% of customers generate 80% of revenue. Or, in economics, it might be used to describe the fact that 20% of the population owns 80% of the wealth.

The Pareto principle can be a useful tool for identifying the most important areas to focus on in order to achieve the greatest results. For example, a business owner might use the Pareto principle to identify the products or services that are generating the most revenue, and then focus on marketing and selling those products or services more aggressively.



🔗 **Ref- What is the Pareto principle**

51. What are left-skewed and right-skewed distributions?

Skewness is a way to describe the symmetry of a distribution.

A left-skewed (Negative Skew) distribution is one in which the left tail is longer than that of the right tail. For this distribution, ***mean < median < mode*.**

Similarly, right-skewed (Positively Skew) distribution is one in which the right tail is longer than the left one. For this distribution, ***mean > median > mode.***

52. If a distribution is skewed to the right and has a median of 20, will the mean be greater than or less than 20?

Ans- In a right-skewed distribution, the mean is greater than the median. This is because the tail of the distribution is on the right, meaning that there are more values greater than the median than there are values less than the median.

Here is a simple example:

Distribution: 10, 15, 20, 25, 30, 35, 40

Median: 25 Mean: 26

This distribution is skewed to the right because there are more values greater than the median (30, 35, 40) than there are values less than the median (10, 15, 20).

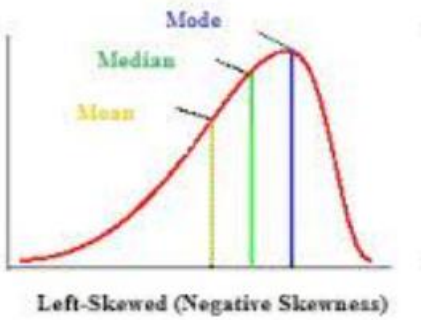
Another way to think about it is that the mean is pulled up by the larger values in the distribution. In a right-skewed distribution, the larger values are on the right side of the distribution, so they pull the mean up.

The🔗 Refore, if a distribution is skewed to the right and has a median of 20, the mean will be greater than 20.

53. Given a left-skewed distribution that has a median of 60, what conclusions can we draw about the mean and the mode of the data?

Ans- In a left-skewed distribution, the mean is typically less than the median. This is because the distribution has a long tail on the left side, which means that there are more values below the median than above it. The mode is also typically less than the median, but not as much as the mean. The median is the middle value in the distribution, so it is not affected by the skew. The mean is pulled down by the long tail on the left side, so it is less than the median. The mode is the most frequent value in the distribution, so it is somewhere near the middle, but it is also pulled down by the skew.

The🔗 Refore, we can conclude that the mean and mode of a left-skewed distribution are both less than the median.



54. Imagine that Jeremy took part in an examination. The test has a mean score of 160, and it has a standard deviation of 15. If Jeremy's z-score is 1.20, what would be his score on the test?

Ans- The z-score formula is z = (x - mu) / sigma, where x is the raw score, mu is the population mean, and sigma is the population standard deviation.

We are given that the population mean is 160 and the population standard deviation is 15. We are also given that Jeremy's z-score is 1.20.

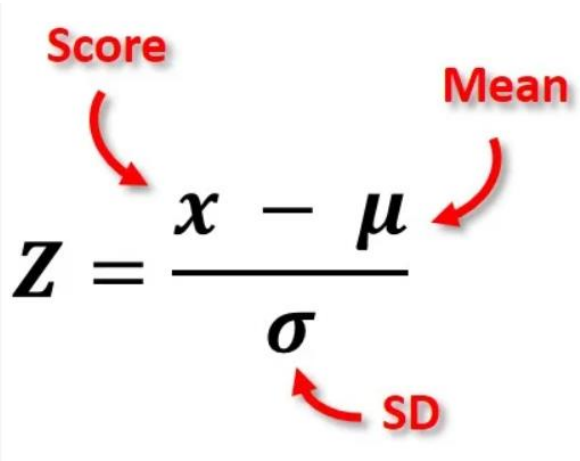
We can use the z-score formula to calculate Jeremy's raw score:

x = mu + sigma * z

x = 160 + 15 * 1.20

x = 186

The 🔗 Refore, Jeremy's score on the test is 186.



🔗 **Ref- [Z-Score & App]**

55. The standard normal curve has a total area to be under one, and it is symmetric around zero. True or False?

Ans- True: The standard normal curve has a total area of 1 and is symmetric around zero.

Explanation:

The standard normal curve is a probability distribution with a mean of 0 and a standard deviation of 1. It is also known as the bell curve because of its bell-shaped shape. The total area under the standard normal curve is equal to 1, which means that there is a 100% probability that a standard normal variable will take on some value.

The standard normal curve is also symmetric around zero. This means that the area under the curve to the left of zero is equal to the area under the curve to the right of zero. In other words, there is a 50% probability that a standard normal variable will be less than zero and a 50% probability that it will be greater than zero.

🔗 **Ref- [Standard Normal Distribution]**

56. Briefly explain the procedure to measure the length of all sharks in the world.

Ans- To measure the length of all sharks in the world, we would need to develop a machine learning model that can accurately estimate the length of a shark from a video or image. This model could be trained on a dataset of labeled images or videos of sharks, where the length of each shark is known.

Once the model is trained, it could be used to measure the length of sharks in real-time, or in videos or images that have already been recorded. This could be done by deploying the model to a mobile app or website, or by integrating it into existing video surveillance systems.

Here is a brief overview of the procedure:

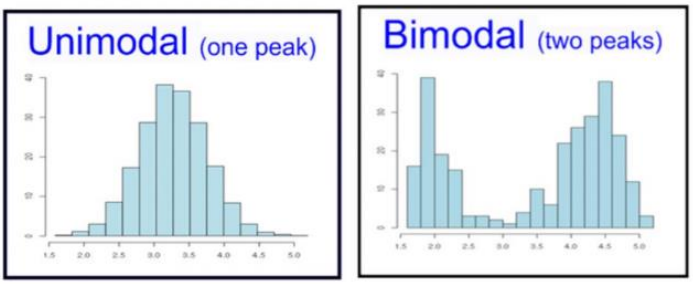
1. Collect a dataset of labeled images or videos of sharks, where the length of each shark is known.
2. Train a machine learning model to estimate the length of a shark from a video or image.
3. Deploy the model to a mobile app, website, or video surveillance system.
4. Use the model to measure the length of sharks in real-time, or in videos or images that have already been recorded.

57. Can you tell me the difference between unimodal bimodal and bell-shaped curves?

Ans- **Unimodal distribution**: A distribution with one peak. This is the most common type of distribution, and it is often symmetrical, meaning that the left and right halves of the distribution are mirror images of each other. Examples of unimodal distributions include the normal distribution, which is also known as the bell-shaped curve.

Bimodal distribution: A distribution with two peaks. This type of distribution is less common than unimodal distributions, and it is often asymmetrical. Examples of bimodal distributions include the distribution of heights of men and women, and the distribution of test scores for students who took the exam twice.

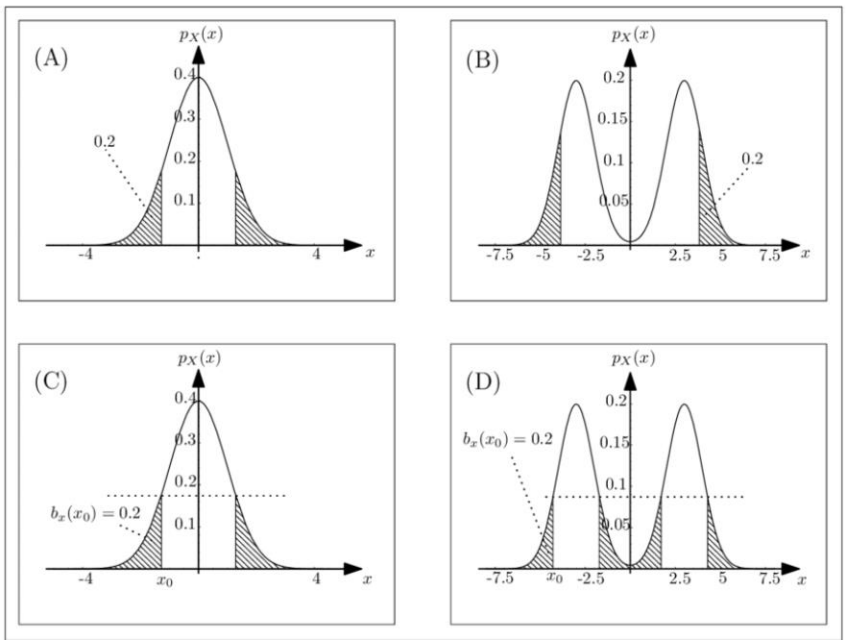
Bell-shaped curve: A symmetrical unimodal distribution. This type of distribution is also known as the normal distribution, and it is characterized by a single peak in the middle, with the values gradually decreasing as you move away from the peak.



58. Does symmetric distribution need to be unimodal?

Ans-

A symmetric distribution does not need to be unimodal. A symmetric distribution is one where the left and right sides of the distribution are mirror images of each other. A unimodal distribution is one where there is a single peak in the distribution.



59. What are some examples of data sets with non-Gaussian distributions?

Ans. When data follows a non-normal distribution, it is frequently non-Gaussian. A non-Gaussian distribution is often seen in many statistics processes. This occurs when data is naturally clustered on one side or the other on a graph. For instance, bacterial growth follows an exponential or non-Gaussian distribution, which is non-normal.

60. What is the Binomial Distribution Formula?

Ans- he Binomial Distribution Formula is:

$$P(X) = {}_nC_x p^x (1-p)^{n-x}$$

where:

- P(X = k) is the probability of getting k successes in n trials
- nCk is the number of ways to choose k successes out of n trials
- pk is the probability of success on each trial
- q is the probability of failure on each trial (q = 1 - p)

Example:

Suppose we flip a coin 10 times. What is the probability of getting 5 heads?

n = 10 (number of trials) k = 5 (number of successes) p = 0.5 (probability of success on each trial) q = 0.5 (probability of failure on each trial)

P(X = 5) = 10C5 * 0.5^5 * 0.5^(10-5) = 252 * 0.03125 = 0.078125

61. What are the criteria that Binomial distributions must meet?

Ans-

Criteria for Binomial Distributions

1. Fixed number of trials

The binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent trials, where each trial has two possible outcomes (success or failure) with the same probability of success.

2. Two mutually exclusive outcomes

In a binomial distribution, there are only two possible outcomes on each trial: success or failure. These outcomes must be mutually exclusive, meaning that only one outcome can occur on each trial.

3. Independent trials

The trials in a binomial distribution must be independent, meaning that the outcome of one trial does not affect the outcome of any other trial.

4. Constant probability of success

The probability of success on each trial in a binomial distribution must be the same. This probability is denoted by the parameter p.

Example

Suppose we flip a coin 10 times. Each flip is a trial, and the two possible outcomes are heads (success) and tails (failure). The probability of success on each flip is the same, which is 0.5. The number of heads we get in 10 flips is a binomial random variable.

62. What are the examples of symmetric distribution?

Ans-

A symmetric distribution is a probability distribution where the left and right halves of the distribution are mirror images of each other. This means that the mean, median, and mode of the distribution are all equal.

Here are some examples of symmetric distributions:

- Normal distribution: The normal distribution is the most common type of symmetric distribution. It is often Referred to as the "bell curve" because of its characteristic shape.
- Uniform distribution: The uniform distribution is a distribution where all values within a certain range are equally likely. This distribution is often used to model random events, such as the flip of a coin or the roll of a die.
- Binomial distribution: The binomial distribution is a distribution that models the number of successes in a fixed number of trials. This distribution is often used to model events such as the number of heads obtained in 10 flips of a coin or the number of defective products in a batch of 100 products.

63. **How to find the mean length of all fishes in the sea?**

Define the confidence level (most common is 95%)
Take a sample of fishes from the sea (to get better results the number of fishes > 30)
Calculate the mean length and standard deviation of the lengths
Calculate t-statistics
Get the confidence interval in which the mean length of all the fishes should be.

64. What are the types of sampling in Statistics?

Ans- There are two main types of sampling in statistics: probability sampling and non-probability sampling.

Probability sampling gives every member of the population an equal chance of being selected for the sample. This makes it the most unbiased sampling method. There are four main types of probability sampling:

- Simple random sampling (SRS): Each member of the population is assigned a unique number, and then random numbers are generated to select the sample.
- Systematic sampling: Every nth member of the population is selected for the sample.
- Stratified sampling: The population is divided into subgroups, or strata, and then a random sample is taken from each stratum. This ensures that the sample is representative of the population in terms of important characteristics, such as age, gender, or region.
- Cluster sampling: The population is divided into clusters, and then a random sample of clusters is selected. All members of the selected clusters are then included in the sample. This method is often used when it is difficult or expensive to obtain a list of all members of the population.

Non-probability sampling does not give every member of the population an equal chance of being selected for the sample. This makes it a more biased sampling method, but it can be useful in some cases, such as when it is difficult or expensive to obtain a random sample. There are four main types of non-probability sampling:

- Convenience sampling: The sample is selected from members of the population who are easy to access. This method is often used in pilot studies or exploratory research.
- Judgment sampling: The sample is selected by the researcher based on their expertise and judgment. This method is often used when the researcher needs to obtain a sample of experts or specialists.
- Snowball sampling: The sample is selected by asking participants to 🔗 Refer other participants who meet the criteria for the study. This method is often used in studies of hidden populations, such as drug users or sex workers.
- Quota sampling: The population is divided into subgroups, or quotas, and then a non-random sample is taken from each quota. This method is often used in market research to ensure that the sample is representative of the population in terms of important characteristics, such as age, gender, and income.

🔗 Ref types of sampling in Statistics

Why is sampling required?

Ans- Sampling is required in machine learning for several reasons:

- Computational efficiency: Machine learning models can be computationally expensive to train, especially on large datasets. Sampling allows us to train models on smaller, more manageable datasets, while still getting accurate results.
- Statistical accuracy: Sampling can actually improve the statistical accuracy of machine learning models. This is because it allows us to avoid overfitting, which is when a model learns the training data too well and cannot generalize to new data.
- Data privacy: In some cases, we may not have access to the entire dataset, or we may need to protect the privacy of the data subjects. Sampling allows us to train models on a subset of the data without compromising privacy.

🔥 **Bonus tip:** It is important to use a sampling method that is representative of the overall population. This means that the sample should have the same distribution of characteristics as the population. If the sample is not representative, the model may be biased and produce inaccurate predictions.

🔗 Ref- Why sampling is required

65. How do you calculate the needed sample size?

Ans- To calculate the needed sample size, you need to consider the following factors:

- The desired confidence level: This is the probability that your sample results will accurately 🔗 Reflect the population results. A higher confidence level will require a larger sample size.
- The margin of error: This is the maximum amount of error you are willing to accept in your sample results. A smaller margin of error will require a larger sample size.
- The population standard deviation: This is a measure of how spread out the population is. A larger population standard deviation will require a larger sample size.

$$\text{Necessary Sample Size} = \frac{(\text{Z-score})2 \times \text{StdDev} \times (1\text{-StdDev})}{(\text{margin of error})2}$$

🔗 Ref- calculate the needed sample size

67. Can you give the difference between stratified sampling and clustering sampling?

Ans- Stratified sampling is a probability sampling technique in which the population is divided into groups, or strata, based on a specific characteristic, such as age, gender, or location. A random sample is then taken from each stratum. This ensures that the sample is representative of the population as a whole, even if the different groups are not equally represented in the population.

Cluster sampling is also a probability sampling technique, but instead of dividing the population into strata, it divides the population into groups, or clusters. A random sample of clusters is then selected, and all members of the selected clusters are included in the sample. This technique is often used when it is difficult or expensive to sample the entire population.

Characteristic	Stratified sampling	Cluster sampling
Population is divided into groups	Yes	Yes
Sample is taken from each group	Yes	No
Sample is representative of the population as a whole	Yes	Yes, but only if the clusters are representative of the population as a whole
Used when	The population is not homogeneous (i.e., different groups are not equally represented in the population)	It is difficult or expensive to sample the entire population

🔥 **Bonus tip:** Stratified sampling is more complex to implement than cluster sampling, but it produces more representative samples. Cluster sampling is less complex to implement, but it can produce less representative samples, especially if the clusters are not representative of the population as a whole.

🔗 Ref-[stratified sampling Vs clustering sampling]

68. Where is inferential statistics used?

Ans- Inferential statistics is used in a wide variety of fields, including:

- Machine learning: Inferential statistics is used to train and evaluate machine learning models. For example, it can be used to test the hypothesis that a particular model is generalizing well to new data.
- Data science: Inferential statistics is used to analyze data and draw conclusions about the underlying population. For example, it can be used to test the hypothesis that there is a significant difference in the average height of men and women.
- Business analytics: Inferential statistics is used to make decisions about products, marketing campaigns, and other business operations. For example, it can be used to test the hypothesis that a new product launch will be successful.
- Scientific research: Inferential statistics is used to test hypotheses about the natural world. For example, it can be used to test the hypothesis that a new drug is effective in treating a particular disease.

🔥 **Bonus tip:** Inferential statistics is particularly useful in situations where we cannot directly measure the entire population of interest. Instead, we take a sample of the population and use inferential statistics to draw conclusions about the entire population.

For example, we cannot directly measure the average height of all men in the world. However, we can take a sample of men and use inferential statistics to estimate the average height of all men.

Inferential statistics is a powerful tool that can be used to make informed decisions in a wide variety of fields.

🔗 Ref- Inferential Statistics

69. What are population and sample in Inferential Statistics, and how are they different?

Ans- Population and sample in inferential statistics

Population: The entire group of individuals, objects, or events that a researcher is interested in studying.

Sample: A subset of the population that is selected for study.

Difference between population and sample:

- Population is the entire group, while sample is a subset of the population.
- Population is usually too large to study in its entirety, so researchers use samples to make inferences about the population.
- Samples should be representative of the population to ensure that the inferences made about the population are accurate.

Example:

A researcher wants to study the average height of adult women in the United States. The population is all adult women in the United States. The sample is a subset of adult women in the United States who are selected for the study.

🔗 **Ref- Population and sample in inferential statistics**

70. What is the relationship between the confidence level and the significance level in statistics?

Ans- The relationship between the confidence level and the significance level in statistics is expressed as:

Confidence level = 1 - Significance level (alpha)

In other words, the confidence level equals one minus the significance level.

The confidence level tells you how sure you can be that your conclusions are correct, while the significance level tells you the probability of rejecting the null hypothesis when it is true (Type I error).

For example, if you set a significance level of 0.05, this means that there is a 5% chance of rejecting the null hypothesis when it is true. The corresponding confidence level would be 1 - 0.05 = 0.95, or 95%.

This means that you can be 95% confident that your conclusions are correct, if you reject the null hypothesis.

🔗 Ref-Confidence Interval

71. What is the difference between Point Estimate and Confidence Interval Estimate?

Ans- A point estimate is a single value that is used to estimate a population parameter. For example, the sample mean is a point estimate of the population mean. A confidence interval estimate is a range of values that is likely to contain the population parameter. For example, a 95% confidence interval estimate for the population mean is a range of values that we are 95% confident contains the true population mean.

Point estimates are easier to calculate and communicate than confidence interval estimates, but they are also less informative. Confidence interval estimates provide more information about the uncertainty associated with the estimate, but they are more difficult to calculate and communicate.

Example:

Suppose we want to estimate the average height of all adults in the United States. We could take a sample of 100 adults and measure their heights. The sample mean height would be a point estimate of the population mean height. However, we know that the sample mean height is not likely to be exactly equal to the population mean height. This is because our sample is just a small subset of the population.

We can use a confidence interval to estimate the range of values that is likely to contain the population mean height. For example, a 95% confidence interval for the population mean height might be 69 inches to 71 inches. This means that we are 95% confident that the true population mean height is somewhere between 69 inches and 71 inches.

🔥 **Bonus tip:** Which one to use?

The choice of whether to use a point estimate or a confidence interval estimate depends on the specific needs of the situation. Point estimates are easier to calculate and communicate, but they are less informative. Confidence interval estimates provide more information about the uncertainty associated with the estimate, but they are more difficult to calculate and communicate.

If we need to make a quick decision and do not need to know the full extent of the uncertainty associated with the estimate, then a point estimate may be sufficient. However, if we need to make a more important decision or need to communicate the uncertainty associated with the estimate, then a confidence interval estimate is more appropriate.

🔗 Ref- Point Estimate and Confidence Interval Estimate

72. What do you understand about biased and unbiased terms?

Ans- Biased and unbiased terms in ML

Biased terms are those that are associated with a particular group or demographic, and can be used in a way that is discriminatory or unfair. For example, the term "blacklisted" is biased because it is associated with negative consequences, such as being denied employment or housing.

Unbiased terms are those that are neutral and can be used to describe any group or demographic without discrimination. For example, instead of saying "blacklisted," you could say "added to a denylist."

🔥 **Bonus tip:** It is important to use unbiased terms in ML to avoid creating models that are discriminatory or unfair. For example, if you are training a model to predict recidivism, and you use biased terms in the training data, the model may learn to discriminate against certain groups of people.

Here are some tips for using unbiased terms in ML:

- Avoid terms that are associated with negative stereotypes or connotations.
- Use neutral terms that can be used to describe any group or demographic.
- Be aware of the context in which you are using terms, and make sure that they are not being used in a discriminatory way.

It is also important to note that there is no single, definitive list of biased and unbiased terms. The meaning of a term can vary depending on the context in which it is used. The 🔗 Refore, it is important to be mindful of the language you are using and to be respectful of all groups of people.

🔗 **Ref- biased and unbiased**

73. How does the width of the confidence interval change with length?

Ans- The width of a confidence interval changes inversely with length. This means that a longer confidence interval will have a narrower width, and a shorter confidence interval will have a wider width. This is because a longer confidence interval is more precise, and a wider confidence interval is less precise.

🔥 **Bonus tip:** The width of a confidence interval can be affected by two factors:

- Sample size: A larger sample size will result in a narrower confidence interval.
- Standard deviation: A higher standard deviation will result in a wider confidence interval.

The standard deviation is a measure of how spread out the data is. If the data is more spread out, then the sample mean is less reliable, and we need a wider confidence interval to account for this uncertainty.

74. What is the meaning of standard error?

Ans- Standard error is a measure of how much the sample mean is likely to vary from the population mean. It is calculated by dividing the standard deviation of the sample by the square root of the sample size.

🔥 **Bonus tip:** Standard error tells us how accurate our sample mean is. It is calculated by taking the standard deviation of the sample and dividing it by the square root of the sample size. The larger the sample size, the smaller the standard error, which means that the sample mean is more likely to be close to the population mean.

75. What is a Sampling Error and how can it be reduced?

Ans- A sampling error is the difference between the true value of a population parameter and the

estimated value of that parameter based on a sample. It is caused by the fact that no sample is perfectly representative of the entire population.

Sampling errors can be reduced by:

- Increasing the sample size. The larger the sample, the less likely it is that it will be unrepresentative of the population.
- Using a good sampling method. A good sampling method is one that gives all members of the population an equal chance of being selected.
- Stratifying the sample. Stratification involves dividing the population into groups (strata) and then taking a random sample from each stratum. This ensures that all groups in the population are represented in the sample.

🔗 Ref- how can it be reduced

76. How do the standard error and the margin of error relate?

Ans-

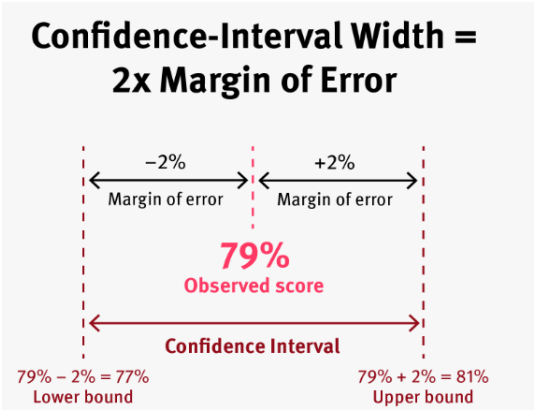
Standard error and margin of error are two closely related concepts in statistics. Both measure the uncertainty of a sample statistic, but they have slightly different interpretations.

Standard error is a measure of the variability of a sample statistic. It tells us how much the sample statistic would vary if we were to repeat the study using new samples from the same population.

Margin of error is a measure of the uncertainty of a sample statistic in relation to the population parameter it is estimating. It tells us how much the sample statistic is likely to differ from the true population parameter.

The margin of error is calculated by multiplying the standard error by a critical value, which depends on the desired confidence level. The higher the confidence level, the larger the margin of error.

Here is a simple analogy to help understand the relationship between standard error and margin of error:



77. What is hypothesis testing?

Ans- Hypothesis testing is a statistical method used to determine whether the relationship between two or more variables is statistically significant. It is a way to test whether a claim about a population is likely to be true.

Hypothesis testing is often used in machine learning to evaluate the performance of a model. For example, you might use hypothesis testing to test whether a new model is more accurate than an existing model.

Here is a short explanation of the steps involved in hypothesis testing:

1. State your null hypothesis and alternative hypothesis. The null hypothesis is the claim that you are trying to disprove. The alternative hypothesis is the claim that you are trying to prove.
2. Collect data. You will need to collect data from a sample of the population that you are interested in.
3. Calculate the test statistic. The test statistic is a measure of how different your sample data is from what would be expected under the null hypothesis.
4. Determine the p-value. The p-value is the probability of obtaining a test statistic as extreme or more extreme than the one you observed, assuming that the null hypothesis is true.
5. Make a decision. If the p-value is less than a certain threshold (usually 0.05), then you reject the null hypothesis and accept the alternative hypothesis. Otherwise, you fail to reject the null hypothesis.

🔗 **Ref- hypothesis testing**

78. What is an alternative hypothesis?

Ans- An alternative hypothesis is a statement that contradicts the null hypothesis. It is the hypothesis that the researcher is trying to prove. In machine learning, the alternative hypothesis is typically a statement that there is a significant relationship between the independent and dependent variables, or that the model performs better than a baseline model.

For example, if the null hypothesis is that there is no relationship between the price of a house and the number of bedrooms, then the alternative hypothesis might be that there is a positive relationship between the two variables. Or, if the null hypothesis is that a new machine learning model performs no better than a random baseline, then the alternative hypothesis might be that the new model performs significantly better than the baseline.

The alternative hypothesis is typically denoted by Ha or H1. It is important to note that the alternative hypothesis is not the same as the research hypothesis. The research hypothesis is the

general statement that the researcher is trying to prove, while the alternative hypothesis is the specific statement that is tested in the statistical test.

🔥 **Bonus tip:** The alternative hypothesis is an important concept in machine learning, as it allows researchers to test their hypotheses about the data and models. By understanding the alternative hypothesis, you can better understand the results of statistical tests and make more informed decisions about your machine learning projects.

🔗 Ref-What is an alternative hypothesis

79. What is the difference between one-tailed and two-tail hypothesis testing?

Ans- One-tailed hypothesis testing is a type of hypothesis testing where the alternative hypothesis specifies a direction for the effect. For example, you might want to test the hypothesis that a new drug is more effective than a placebo in reducing cholesterol levels. In this case, your alternative hypothesis would be that the mean cholesterol level of the group taking the new drug is greater than the mean cholesterol level of the group taking the placebo.

Two-tailed hypothesis testing is a type of hypothesis testing where the alternative hypothesis does not specify a direction for the effect. For example, you might want to test the hypothesis that a new drug is effective in reducing cholesterol levels. In this case, your alternative hypothesis would be that the mean cholesterol level of the group taking the new drug is different from the mean cholesterol level of the group taking the placebo.

🔥 **Bonus tip:** One-tailed hypothesis tests are more powerful than two-tailed hypothesis tests, but they are also more likely to produce type I errors (false positives). The 🔗 Refore, one-tailed hypothesis tests should only be used when you have a strong prior belief about the direction of the effect.

🔗 Ref- One-tail & two -tail test

80. What is one sample t-test?

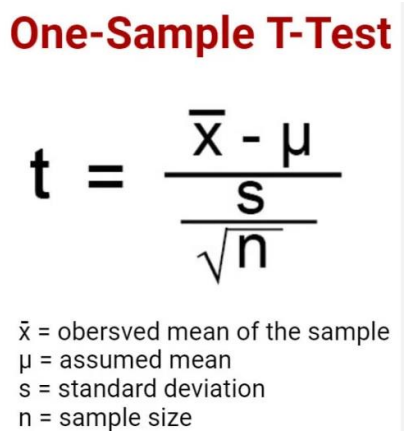
Ans- A one-sample t-test is a statistical hypothesis test used to determine whether the mean of a sample is significantly different from a known population mean. It is a parametric test, which means that it assumes that the data is normally distributed.

To perform a one-sample t-test, we first need to state our null and alternative hypotheses. The null hypothesis is typically that the sample mean is equal to the known population mean. The alternative hypothesis is that the sample mean is different from the known population mean.

Once we have stated our hypotheses, we can calculate the t-statistic. The t-statistic is a measure of how far the sample mean is from the known population mean, in units of the standard error.

To calculate the t-statistic, we use the following formula:

$$t = (\bar{x} - \mu) / (s / \sqrt{n})$$



where:

- $\bar{x}$ is the sample mean
- μ is the known population mean
- s is the sample standard deviation
- n is the sample size

We can then use the t-statistic to look up the p-value in a t-table. The p-value is the probability of obtaining a t-statistic as extreme or more extreme than the one we calculated, assuming that the null hypothesis is true.

If the p-value is less than our significance level (α), we reject the null hypothesis and conclude that there is a statistically significant difference between the sample mean and the known population mean.

🔗 **Ref-T-test**

81. What is the meaning of degrees of freedom (DF) in statistics?

Ans- Degrees of freedom (DF) in statistics is the number of independent pieces of information that go into calculating a statistic. It is often calculated as the sample size minus the number of parameters estimated. For example, if you are calculating the sample mean, the DF is the sample size minus one.

Deeper explanation:

To understand the meaning of DF, it is helpful to think about what it means for a piece of information to be independent. Two pieces of information are independent if they do not contain any common information. For example, if you flip a coin twice, the results of the two flips are independent. However, if you know that the first flip was heads, then the second flip is not independent, because you know that the probability of tails is now higher.

In statistics, we often use samples to estimate population parameters. For example, we might use a sample of students to estimate the average height of all students. However, we know that the sample mean will not be exactly equal to the population mean. Instead, it will be an estimate. The DF tells us how much uncertainty there is in our estimate.

🔥 **Bonus tip:** The more DF we have, the more confident we can be in our estimate. This is because more DF means that we have more independent pieces of information to go on. For example, if we have a sample of 100 students, we are more confident in our estimate of the average height than if we only have a sample of 10 students.

$$df = N - 1$$

🔗 Ref- Degree of freedom

82. What is the p-value in hypothesis testing?

A p-value is a number that describes the probability of finding the observed or more extreme results when the null hypothesis (H0) is True. P-values are used in hypothesis testing to help decide whether to reject the null hypothesis or not. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

83. How can you calculate the p-value?

Ans- A p-value is the probability of obtaining a result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. It is used in hypothesis testing to determine whether the observed results are statistically significant.

To calculate the p-value, you need to first determine the test statistic. The test statistic is a measure of how different the observed results are from the expected results. Once you have calculated the test statistic, you need to look up the p-value in a p-value table or use statistical software.

The p-value is typically compared to a significance level, which is denoted by alpha. The significance level is the probability of rejecting the null hypothesis when it is true. It is typically set to 0.05, but it can be adjusted depending on the situation.

If the p-value is less than the significance level, then the null hypothesis is rejected. This means that there is enough evidence to suggest that the null hypothesis is false and that the alternative hypothesis is true.

🔗 **Ref-P-value**

84. If there is a 30 percent probability that you will see a supercar in any 20-minute time interval, what is the probability that you see at least one supercar in the period of an hour (60 minutes)?

Ans- To find the probability of seeing at least one supercar in a 60-minute period when there's a 30 percent probability of seeing one in a 20-minute interval, we can use the complement rule.

The complement rule states that the probability of an event occurring is equal to 1 minus the probability of the event not occurring.

In this case, the probability of not seeing a supercar in a 20-minute interval is 1 - 0.30 = 0.70.

Now, let's calculate the probability of not seeing a supercar in three consecutive 20-minute intervals (60 minutes in total):

0.70 (first 20 minutes) * 0.70 (second 20 minutes) * 0.70 (third 20 minutes) = 0.343.

So, the probability of not seeing a supercar in a 60-minute period is 0.343.

To find the probability of seeing at least one supercar in a 60-minute period, you subtract this probability from 1:

1 - 0.343 = 0.657.

So, there's a 65.7 percent probability of seeing at least one supercar in a 60-minute period.

Hypothesis testing is a type of statistical inference that uses data from a sample to conclude about the population data.
Before performing the testing, an assumption is made about the population parameter. This assumption is called the null hypothesis and is denoted by H0. An alternative hypothesis (denoted Ha), which is the logical opposite of the null hypothesis, is then defined.
The hypothesis testing procedure involves using sample data to determine whether or not H0 should be rejected. The acceptance of the alternative hypothesis (Ha) follows the rejection of the null hypothesis (H0).

85. How would you describe a ‘p-value’?

Ans- A p-value is a measure of the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. It is a useful tool for machine learning practitioners to assess the statistical significance of their results.

86. What is the difference between type I vs type II errors?

Ans- Type I and type II errors are two types of errors that can occur in hypothesis testing.

Type I error is when we reject a true null hypothesis. This is also known as a false positive. For example, a type I error would occur if we concluded that a drug was effective in treating a disease when it was actually not.

Type II error is when we fail to reject a false null hypothesis. This is also known as a false negative. For example, a type II error would occur if we concluded that a drug was not effective in treating a disease when it was actually effective

🔥 **Bonus tip:** There are a number of things that can be done to reduce the risk of making type I and type II errors, including:

- Using a larger sample size
- Using a more powerful statistical test
- Setting a lower significance level for type I errors
- Increasing the power of the test against type II errors

It is important to balance the risk of making type I and type II errors when designing a hypothesis test. The optimal balance will depend on the specific context of the test.

87. When should you use a t-test vs a z-test?

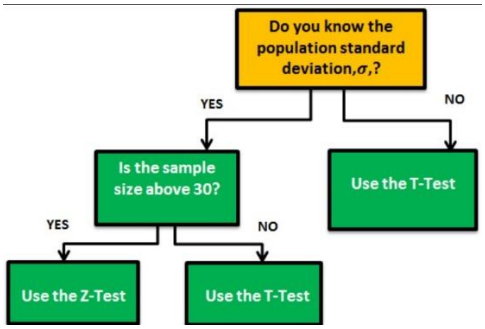
Ans- A t-test is a parametric test used to compare the means of two groups when the population standard deviation is unknown. A z-test is also a parametric test, but it is used when the population standard deviation is known.

In general, you should use a t-test if:

- The population standard deviation is unknown.
- The sample size is less than 30.

You should use a z-test if:

- The population standard deviation is known.
- The sample size is greater than 30.



🔗 **Ref-Z Vs T**

88. What is the difference between the f test and anova test?

Ans- The F-test is a statistical test that compares two variances. It is often used to test the null hypothesis that two groups have the same variance. The ANOVA test is a statistical test that compares the means of three or more groups. It is often used to test the null hypothesis that all of the groups have the same mean.

In other words, the F-test is used to test for homogeneity of variance, while the ANOVA test is used to test for homogeneity of means.

Here is a table summarizing the key differences between the F-test and the ANOVA test:

Characteristic	F-test	ANOVA test
Purpose	Tests for homogeneity of variance	Tests for homogeneity of means
Number of groups	Two or more	Three or more
Null hypothesis	Two groups have the same variance	All groups have the same mean
Test statistic	F-statistic	F-statistic
Degrees of freedom	Between groups and within groups	Between groups and within groups

🔗 Ref-F Vs ANOVA

89. What is Resampling and what are the common methods of resampling?

Ans- Resampling is a statistical technique that involves repeatedly sampling from a given dataset to create new datasets. It can be used to estimate the variability of a statistic, such as the mean or median, or to create new training data for machine learning models.

Common Resampling Methods:

- **Bootstrapping**: Bootstrapping involves creating new datasets by randomly sampling with replacement from the original dataset. This means that some data points may be included multiple times in a single bootstrap sample, while others may not be included at all.
- **Cross-Validation**: Cross-validation involves splitting the original dataset into multiple folds. A model is then trained on each fold using the data from the other folds as validation data. This process is repeated for all folds, and the average performance of the model on the validation sets is used as an estimate of its generalization performance.
- **Permutation Testing**: Permutation testing involves randomly shuffling the labels of the data points in the original dataset. A statistic is then calculated for the shuffled dataset, and this process is repeated many times. The p-value of the test is then calculated as the proportion of shuffled datasets for which the statistic is greater than or equal to the statistic calculated on the original dataset.

> 🔥 **Bonus tip:** Resampling is a powerful statistical technique that can be used in a variety of applications. It is a valuable tool for machine learning practitioners, as it can be used to estimate the variability of statistics, create new training data, and evaluate the performance of machine learning models.

🔗 Ref- Resampling

90. What is the proportion of confidence intervals that will not contain the population parameter?

Ans- A confidence interval is a range of values that is likely to contain the true population parameter. The confidence level is the probability that the confidence interval will actually contain the population parameter. For example, a 95% confidence interval means that there is a 95% probability that the interval will contain the true population parameter.

It is important to note that a confidence interval is a statistical estimate, and it is not guaranteed to contain the population parameter. The proportion of confidence intervals that will not contain the population parameter is equal to the complement of the confidence level. For example, a 95% confidence interval will not contain the population parameter 5% of the time.

🔥 **Bonus tip:** The proportion of confidence intervals that will not contain the population parameter is equal to the complement of the confidence level. For example, a 95% confidence interval will not contain the population parameter 5% of the time.

91. What is a confounding variable?

A confounding variable in statistics is an 'extra' or 'third' variable that is associated with both the dependent variable and the independent variable, and it can give a wrong estimate that provides useless results. For example, if we are studying the effect of weight gain, then lack of workout will be the independent variable, and weight gain will be the dependent variable. In this case, the amount of food consumption can be the confounding variable as it will mask or distort the effect of other variables in the study. The effect of weather can be another confounding variable that may later the experiment design.

92. What are the steps we should take in hypothesis testing?
Ans.

1. State the null hypothesis
2. State the alternate hypothesis
3. Which test and test statistic to be performed
4. Collect Data
5. Calculate the test statistic
6. Construct Acceptance / Rejection regions
7. Based on steps 5 and 6, draw a conclusion about H0

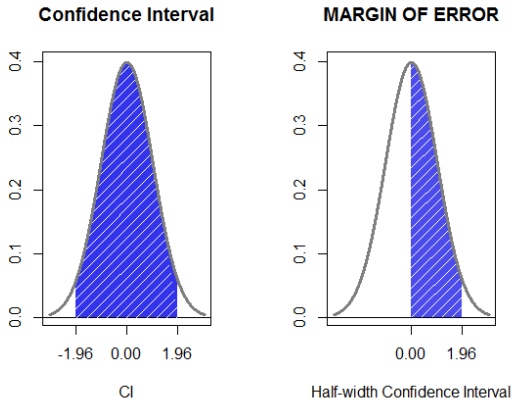
93. What is the relationship between standard error and the margin of error?

Ans- The margin of error is used to construct confidence intervals, which are ranges of values that are likely to contain the true population parameter. The confidence level of a confidence interval is the probability that the true population parameter is within the interval.

The relationship between standard error and margin of error is as follows:

```
Margin of error = Standard error * Critical value
```

The critical value depends on the desired confidence level and the type of statistical distribution that is being used. For example, the critical value for a 95% confidence interval using the normal distribution is 1.96.



94. How would you describe what a 'p-value' is to a non-technical person or in a layman term?

The best way to describe the p-value in simple terms is with an example. In practice, if the p-value is less than the alpha, say of 0.05, then we're saying that there's a probability of less than 5% that the result could have happened by chance. Similarly, a p-value of 0.05 is the same as saying "5% of the time, we would see this by chance."

95. What does interpolation and extrapolation mean? Which is generally more accurate?

Interpolation is a prediction made using inputs that lie within the set of observed values. Extrapolation is when a prediction is made using an input that's outside the set of observed values.

Generally, interpolations are more accurate.

96. What is an inlier?

An inlier is a data observation that lies within the rest of the dataset and is unusual or an error. Since it lies in the dataset, it is typically harder to identify than an outlier

97. You roll a biassed coin (p(head)=0.8) five times. What's the probability of getting three or more heads?

and requires external data to identify them. Should you identify any inliers, you can simply remove them from the dataset to address them.

Ans- the probability of getting 3 or more heads when rolling a biased coin with probability of head 0.8 five times:

```
Probability of 5 heads = 0.8^5 = 0.32768
Probability of 4 heads = 0.8^4 * 0.2^1 * 5C4 = 0.21504
Probability of 3 heads = 0.8^3 * 0.2^2 * 5C3 = 0.1024
```

The 🔗 Refore, the probability of getting 3 or more heads is 0.32768 + 0.21504 + 0.1024 = 0.64512

98. Infection rates at a hospital above a 1 infection per 100 person-days at risk are considered high. A hospital had 10 infections over the last 1787

person-days at risk. Give the p-value of the correct one-sided test of whether the hospital is below the standard

Ans-the standard of 1 infection per 100 person-days at risk, we can use a one-sided proportion test. The null hypothesis is that the infection rate is greater than or equal to the standard, and the alternative hypothesis is that the infection rate is less than the standard.

To calculate the p-value, we first need to calculate the z-statistic:

```
z = (p - p_0) / sqrt(p_0(1 - p_0) / n)
```

where:

- p is the observed proportion of infections (10/1787)
- p_0 is the standard proportion of infections (1/100)
- n is the sample size (1787)

```
z = (0.0056 - 0.01) / sqrt(0.01(1 - 0.01) / 1787)
= -0.56
```

We can then look up the p-value of this z-statistic in a standard normal table. The p-value is 0.2877.

This p-value tells us that there is a 28.77% chance of observing a sample proportion of infections this low or lower if the true infection rate is equal to the standard. Since this p-value is greater than 0.05, we fail to reject the null hypothesis. There is not enough evidence to conclude that the hospital's infection rate is below the standard.

.

99. In a population of interest, a sample of 9 men yielded a sample average brain volume of 1,100cc and a standard deviation of 30cc. What is a 95% Student’s T confidence interval for the mean brain volume in this new population?

Ans- In a population of interest, a sample of 9 men yielded a sample average brain volume of 1,100cc and a standard deviation of 30cc. What is a 95% Student's T confidence interval for the mean brain volume in this new population?

Bard: To find a 95% Student's T confidence interval for the mean brain volume, we can use the following formula:

$$(\bar{x} - t_{\alpha/2} * s / \sqrt{n}, \bar{x} + t_{\alpha/2} * s / \sqrt{n})$$

where:

- $\bar{x}$ is the sample mean (1,100cc in this case)
- s is the sample standard deviation (30cc in this case)
- n is the sample size (9 in this case)
- $t_{\alpha/2}$ is the critical value of the Student's T distribution for the desired confidence level and degrees of freedom (df = n - 1 = 8 in this case)

We can find the critical value of the Student's T distribution using a statistical table or calculator. For a 95% confidence level and 8 degrees of freedom, the critical value is `2.262`.

Plugging in the values, we get the following confidence interval:

```
(1,100cc - 2.262 * 30cc / √9, 1,100cc + 2.262 * 30cc / √9)
= (1,077cc, 1,123cc)
```

The 🔗 Refore, we can be 95% confident that the true mean brain volume in this new population is between 1,077cc and 1,123cc.

100. What Chi-square test?

A statistical method is used to find the difference or correlation between the observed and expected categorical variables in the dataset.

Example: A food delivery company wants to find the relationship between gender, location and food choices of people in India.

It is used to determine whether the difference between 2 categorical variables is:

- Due to chance or
- Due to relationship

101. What is the ANOVA test?

Ans- ANOVA, or Analysis of Variance, is a statistical test used to compare the means of two or more groups. It is a parametric test, meaning that it assumes that the data is normally distributed. ANOVA is often used in machine learning to select features and to evaluate the performance of models.

To perform an ANOVA test, we first need to define the null hypothesis and the alternative hypothesis. The null hypothesis is that the means of all the groups are equal. The alternative hypothesis is that at least one group mean is different from the others.

Next, we calculate the F-statistic, which is a measure of the variation between the groups relative to the variation within the groups. If the F-statistic is greater than a critical value, then we reject the null hypothesis and conclude that there is a significant difference between the group means.

🔥 **Bonus tip:** ANOVA is a powerful tool that can be used in a variety of machine learning tasks. It is a good idea for machine learning practitioners to be familiar with ANOVA and how to use it.
🔗 Ref- ANOVA

102. How to calculate p-value using a manual method?

Ans- To calculate a p-value using a manual method, you can follow these steps:

1. State the null and alternative hypotheses. The null hypothesis is the assumption that you are trying to disprove, while the alternative hypothesis is the assumption that you are trying to support.

2. Calculate the test statistic. This is a measure of how different your sample data is from the null hypothesis. The specific test statistic that you use will depend on the type of statistical test that you are performing.
3. Find the p-value for the test statistic. This can be done by using a statistical table, such as a t-distribution table or a chi-squared distribution table. The p-value is the probability of obtaining a test statistic as extreme or more extreme than the one you observed, assuming that the null hypothesis is true.
4. Interpret the p-value. If the p-value is less than a certain significance level (typically 0.05), then you can reject the null hypothesis and conclude that the alternative hypothesis is more likely to be true.

Example:

Suppose that you are interested in testing whether the mean height of male adults in the United States is different from 70 inches. You randomly sample 100 male adults and find that the mean height of the sample is 71 inches.

Null hypothesis: H0: μ = 70 inches Alternative hypothesis: H1: μ ≠ 70 inches

Test statistic:

$$t = (\bar{x} - \mu) / (s / \sqrt{n})$$

where:

- $\bar{x}$ is the sample mean (71 inches)
- μ is the population mean (70 inches)
- s is the sample standard deviation
- n is the sample size (100 adults)

Calculating the p-value:

The p-value can be found using a t-distribution table with 99 degrees of freedom (n - 1). The t-statistic for this example is 2.87, which corresponds to a p-value of 0.005.

Interpreting the p-value:

Since the p-value is less than the significance level of 0.05, we can reject the null hypothesis and conclude that the mean height of male adults in the United States is significantly different from 70 inches.

103. What do we mean by – making a decision based on comparing p-value with significance level?

Ans- Making a decision based on comparing a p-value with a significance level is a statistical method for determining whether there is enough evidence to reject the null hypothesis and conclude that there is a real effect.

The p-value is the probability of obtaining the observed results, assuming that the null hypothesis is true. The significance level is the maximum probability of a Type I error, which is the error of rejecting the null hypothesis when it is true.

To make a decision, we compare the p-value to the significance level. If the p-value is less than or equal to the significance level, we reject the null hypothesis and conclude that there is a real effect. If the p-value is greater than the significance level, we do not reject the null hypothesis and conclude that there is not enough evidence to say that there is a real effect.

It is important to note that a statistically significant result does not necessarily mean that the observed effect is practically significant.

104.What is the goal of A/B testing?

Ans- The goal of A/B testing is to compare two or more variations of a webpage, app, or marketing campaign to determine which one performs better in terms of a specific outcome or metric. A/B testing is commonly used in the field of data-driven decision-making, especially in marketing and product development, to optimize user experiences, increase conversions, and ultimately improve business results.

105. What is the difference between a box plot and a histogram?

Ans- A box plot and a histogram are both data visualization tools used to summarize the distribution of a dataset. However, there are some key differences between the two.

A box plot shows the five-number summary of a dataset: the minimum, median, maximum, first quartile, and third quartile. The box represents the middle 50% of the data, with the line in the middle representing the median. The whiskers extend to the minimum and maximum values, and any outliers are shown as individual dots.

A histogram, on the other hand, shows the frequency distribution of a dataset. This means that it shows how many data points fall into each bin of values. The bins are typically chosen to be of equal width, and the height of each bar represents the number of data points in that bin.

Here is a table summarizing the key differences between box plots and histograms:

🔥 **Bonus tip:**

Feature	Box Plot	Histogram
Type of plot	Categorical	Quantitative
Summary statistic	Five-number summary	Frequency distribution
Outliers	Shows outliers	Does not show outliers

Useful for	Comparing distributions of different datasets	Understanding the distribution of a single dataset

🔗 Ref-difference between a box plot and a histogram

106. A jar has 1000 coins, of which 999 are fair and 1 is double headed. Pick a coin at random, and toss it 10 times. Given that you see 10 heads, what is the probability that the next toss of that coin is also a head?

Ans- The probability that the next toss of the coin is also a head is approximately 0.753.

Explanation:

There are two ways to get 10 heads in 10 tosses:

1. Choose the double-headed coin and toss it 10 times. This will always result in 10 heads.
2. Choose a fair coin and toss it 10 times. This will result in 10 heads with probability 1/1024.

Since the double-headed coin is much more likely to produce 10 heads, it is more likely that the coin you chose was the double-headed coin. The🔗 Refore, the probability that the next toss of the coin is also a head is approximately 0.753.

This can be calculated more precisely using Bayes' theorem:

P(heads next toss | 10 heads in 10 tosses) = P(10 heads in 10 tosses | double-headed coin) * P(double-headed coin) / (P(10 heads in 10 tosses | double-headed coin) * P(double-headed coin) + P(10 heads in 10 tosses | fair coin) * P(fair coin))

= 1 * 1/1000 / (1 * 1/1000 + 1/1024 * 999/1000)

= 0.753

107. What is a confidence interval and how do you interpret it?

Ans- A confidence interval is a range of values that is likely to contain the true population parameter of interest. It is calculated from a sample of data, and the confidence level indicates the probability that the interval contains the true parameter.

For example, a 95% confidence interval for the mean height of all adults in the United States might be [68 inches, 70 inches]. This means that we are 95% confident that the true mean height of all adults in the United States is between 68 inches and 70 inches.

Confidence intervals can be used to estimate a variety of population parameters, such as the mean, median, proportion, and standard deviation. They can also be used to compare two or more population parameters.

To interpret a confidence interval, we simply need to state the confidence level and the range of values. For example, we could say:

We are 95% confident that the true mean height of all adults in the United States is between 68 inches and 70 inches.

This means that if we were to take many different random samples of adults in the United States and calculate a 95% confidence interval for the mean height each time, we would expect 95% of those intervals to contain the true mean height of all adults in the United States.

🔗 Ref- What is a confidence interval

108. How do you stay up-to-date with the new and upcoming concepts in statistics?

Ans- It is important to stay up-to-date with the latest developments in statistics, as it is the foundation of machine learning. Here are some ways I do that:

- **Follow statistics blogs and newsletters**. There are many great resources available online, such as Towards Data Science, KDnuggets, and StatQuest. These resources provide articles on a variety of statistical topics, including new and upcoming concepts.
- **Attend statistics conferences and workshops.** This is a great way to learn about the latest research in statistics and to network with other experts in the field.
- **Read statistics textbooks and research papers.** While this can be more time-consuming, it is a great way to get a deeper understanding of new and upcoming statistical concepts.

109. What is correlation?

Ans- Correlation is a statistical measure of the degree to which two variables are related. It can be positive, negative, or zero. A positive correlation means that the two variables tend to move in the same direction. A negative correlation means that the two variables tend to move in opposite directions. A zero correlation means that there is no relationship between the two variables.

Correlation is often used in machine learning to identify features that are important for predicting a target variable. For example, if we are trying to predict the price of a house, we might look at features such as the square footage, the number of bedrooms, and the location of the house. If the correlation between a feature and the price of the house is high, then that feature is likely to be important for predicting the price.

Correlation can also be used to detect outliers in data. An outlier is a data point that is significantly different from the other data points. Outliers can be caused by errors in data collection or by rare events. By identifying outliers, we can clean our data and improve the performance of our machine learning models.

🔗 **Ref- What is correlation**

110. What types of variables are used for Pearson’s correlation coefficient?

Ans- Pearson's correlation coefficient is a statistical measure of the linear relationship between two quantitative variables. It is calculated using the following formula:

```
r = (cov(x, y)) / (sqrt(var(x)) * sqrt(var(y)))
```

where:

- r is the Pearson correlation coefficient
- cov(x, y) is the covariance of x and y
- var(x) is the variance of x
- var(y) is the variance of y

Types of variables used for Pearson's correlation coefficient:

Both variables must be quantitative, meaning that they can be measured on a numerical scale. Additionally, both variables should be normally distributed.

Example:

Suppose we want to measure the correlation between height and weight in a group of people. Both height and weight are quantitative variables, and they are both normally distributed. The 🔗 Refore, we can use Pearson's correlation coefficient to measure the linear relationship between these two variables.

111.In an observation, there is a high correlation between the time a person sleeps and the amount of productive work he does. What can be inferred from this?

Ans- A high correlation between sleep time and productivity means that people who sleep more are more likely to be productive. This could be due to a number of factors, including:

- Improved cognitive function: Sleep is essential for cognitive function, including attention, memory, and problem-solving. People who are well-rested are better able to focus and think clearly, which can lead to increased productivity.
- Reduced stress: Sleep helps to reduce stress levels. People who are stressed are more likely to be distracted and make mistakes, which can impair productivity.
- Improved mood: Sleep also helps to improve mood. People who are well-rested are more likely to be happy and motivated, which can lead to increased productivity.

However, it is important to note that correlation does not equal causation. Just because there is a high correlation between sleep time and productivity does not mean that sleep is the only factor that affects productivity. Other factors, such as motivation, skills, and resources, also play a role.

112. What is the meaning of covariance?

Ans-

- Covariance is a measure of the linear relationship between two random variables. This means that it only measures how much the two variables change together in a linear way. It does not measure how much they may change together in a nonlinear way.
- Covariance is calculated by taking the average of the products of the deviations of the two variables from their means. This means that it takes into account the size and direction of the deviations of the two variables from their means.
- Covariance can be positive, negative, or zero. A positive covariance indicates that the two variables tend to move in the same direction, while a negative covariance indicates that they tend to move in opposite directions. A covariance of zero indicates that there is no linear relationship between the two variables.

🔥 **Bonus tip:** Applications of covariance:

Covariance is used in a variety of applications, including:

- Machine learning: Covariance is used in many machine learning algorithms, such as linear regression and support vector machines.
- Risk management: Covariance is used in risk management to measure the risk of a portfolio of assets.
- Financial analysis: Covariance is used in financial analysis to measure the relationship between different stocks and market indices.

🔗 Ref-[What is the meaning of covariance]

113. What does autocorrelation mean?

Ans- Autocorrelation is a statistical measure of the correlation between a time series and its lagged values. It is a way of measuring how similar a time series is to itself at different points in time. Autocorrelation can be either positive or negative, and it can be used to identify patterns in time series data.

Example: If the autocorrelation of a time series is high, it means that the values of the time series tend to be similar to the values that came before them. This could be because the time series is following a trend, or because it is being influenced by some other factor that is also changing over time.

How to calculate autocorrelation:

There are a number of ways to calculate autocorrelation, but the most common method is to use the following formula:

$$R_p = \frac{\frac{1}{n-p}\sum\left[\left(V_i - \bar{V}\right)\cdot\left(V_{i+p} - \bar{V}\right)\right]}{\frac{1}{n}\sum\left(V_i - \bar{V}\right)^2}$$

where:

V_i = Historical value in period i

V_{i+p} = Historical value in period i+p

$\bar{V}$ = Mean historical value

n = Number of historical values

p = Number of periods per season

The lagged time series is simply the time series shifted back by a certain number of periods. For example, if we want to calculate the autocorrelation of a time series at lag 1, we would correlate the time series with itself shifted back by one period.

🔥 **Bonus tip:** Autocorrelation is a statistical measure of the correlation between a time series and its lagged values. It is a way of measuring how similar a time series is to itself at different points in time. Autocorrelation can be either positive or negative, and it can be used to identify patterns in time series data.

Autocorrelation is an important concept in time series analysis, and it has a number of applications in machine learning. For example, autocorrelation can be used to feature engineer time series data, to build forecasting models, and to detect anomalies.

🔗 **Ref- [Auto-Correlation]**

114. What types of variables are used for Pearson's correlation coefficient?

Ans- Pearson's correlation coefficient is a statistical measure of the linear relationship between two quantitative variables. It is calculated using the following formula:

```
r = (cov(x, y)) / (sqrt(var(x)) * sqrt(var(y)))
```

where:

- r is the Pearson correlation coefficient
- cov(x, y) is the covariance of x and y
- var(x) is the variance of x

- var(y) is the variance of y

Types of variables used for Pearson's correlation coefficient:

Both variables must be quantitative, meaning that they can be measured on a numerical scale. Additionally, both variables should be normally distributed.

Example:

Suppose we want to measure the correlation between height and weight in a group of people. Both height and weight are quantitative variables, and they are both normally distributed. The 🔗 Refore, we can use Pearson's correlation coefficient to measure the linear relationship between these two variables.

115. How will you determine the test for the continuous data?

Ans- To determine the test for continuous data, I would first need to understand the specific research question or hypothesis I am trying to answer. Once I have a good understanding of the question, I can start to consider the different statistical tests that are available for continuous data.

Some common statistical tests for continuous data include:

- t-test: This test is used to compare the means of two independent groups.
- ANOVA: This test is used to compare the means of three or more independent groups.
- correlation: This test is used to measure the strength and direction of the relationship between two continuous variables.
- regression: This test is used to predict the value of one continuous variable based on the value of another continuous variable.

The specific test that I choose will depend on the nature of my research question and the characteristics of my data. For example, if I am trying to compare the mean height of men and women, I would use a t-test. If I am trying to compare the mean height of men and women from three different countries, I would use an ANOVA test.

🔗 Ref-Test for countinous data

116. What can be the reason for non normality of the data?
Ans- There are many reasons why data can be non-normally distributed, but some of the most common include:

- The underlying distribution is non-normal. This means that the data is naturally distributed in a different way, such as a Poisson distribution or a negative binomial distribution.
- Outliers or mixed distributions are present. Outliers are data points that fall far outside the rest of the data, and mixed distributions are distributions that are made up of two or more different distributions.
- A low discrimination gauge is used. A discrimination gauge is a measure of how well a variable can distinguish between different groups. If the discrimination gauge is low, it can lead to non-normally distributed data.
- Skewness is present in the data. Skewness is a measure of how asymmetrical a distribution is. If the distribution is skewed, it can lead to non-normally distributed data.
- You have a large sample size. As the sample size increases, the data is more likely to be non-normally distributed.

🔥 Bonus tip: If you are working with non-normal data, there are a few things you can do:

- Transform the data. There are a number of statistical transformations that can be used to make non-normal data more normal. One common transformation is the log transformation.
- Use non-parametric tests. Non-parametric tests are statistical tests that do not require the data to be normally distributed. Some common non-parametric tests include the Wilcoxon rank-sum test and the Kruskal-Wallis test.
- Use robust statistical methods. Robust statistical methods are methods that are not sensitive to outliers or non-normality in the data.

It is important to note that there is no one-size-fits-all solution to dealing with non-normal data.

🔗 Ref- Non normality of data

116. why is there no such thing like 3 samples t- test? why t-test failed with 3 samples

Ans- There are two main reasons why a t-test cannot be used to compare the means of more than two groups. First, the t-test statistic is calculated by comparing the difference in the means of the two groups to the standard error of the difference. The standard error of the difference is calculated using the variances of the two groups. When there are more than two groups, it is not clear which variances to use.

Second, the t-test statistic is compared to a t-distribution to determine the p-value. The t-distribution is a theoretical distribution that is based on the number of degrees of freedom. The degrees of freedom for a t-test is the number of samples minus two. When there are more than two groups, it is not clear how to calculate the degrees of freedom.

If a t-test is used to compare the means of more than two groups, the results will be inaccurate. The p-value will be too small, which will lead to an increased risk of Type I errors (false positives).
