## **Bilal Hameed**

## **Data Scientist @ CareCloud**

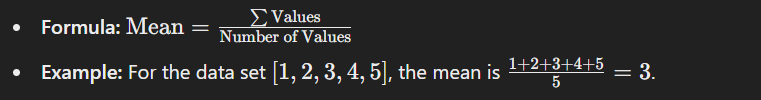
## **Machine Learning Interview Notes**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

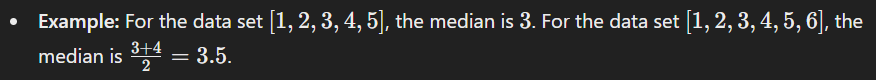
## **Explain the difference between mean, median, and mode. When would you use each?**

The mean, median, and mode are three different measures of central tendency used in statistics to describe the center point of a data set.

## **Mean**

* **Definition:** The mean (or average) is the sum of all the values in a data set divided by the number of values.
* **Use Case:** Best used when you want an overall average and the data does not have extreme outliers.

## **Median**

* **Definition:** The median is the middle value in a data set when the values are arranged in ascending or descending order. If there is an even number of values, the median is the average of the two middle values.
* **Calculation Steps:**
  1. Arrange the data in order.
  2. Find the middle value.
* **Use Case:** Best used when you need the central point of data and want to minimize the effect of outliers.

## **Mode**

* **Definition:** The mode is the value that appears most frequently in a data set. A data set may have one mode, more than one mode, or no mode at all.
* **Example:** For the data set [1,2,2,3,4], the mode is 2 because it appears most frequently. For the data set [1,1,2,3,3], the modes are 1 and 3 (bimodal).
* **Use Case:** Best used when you want to know the most common value(s) in the data set.

## **Summary**

* **Mean:** Sum of values divided by the number of values.
* **Median:** Middle value when data is ordered.
* **Mode:** Most frequent value(s) in the data set.

Each measure provides different insights and is useful in different scenarios depending on the nature of the data and the specific requirements of the analysis.

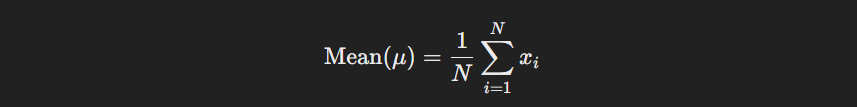
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## **How do you calculate the variance and standard deviation of a dataset?**

To calculate the variance and standard deviation of a dataset, follow these steps:

### 1. **Calculate the Mean (Average):**

* Sum all the data points.
* Divide the sum by the number of data points.



Where *N* is the number of data points and *xi​* represents each data point.

### 2. **Calculate Each Deviation from the Mean:**

* Subtract the mean from each data point to find the deviation of each point from the mean.

Deviation=*xi−μ*

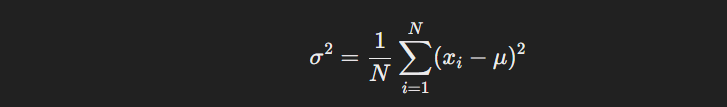
### 3. **Square Each Deviation:**

* Square each of the deviations calculated in step 2 to remove negative values.

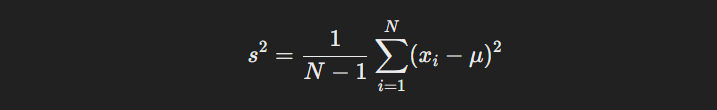
(Deviation)^2=(xi−μ)^2

### 4. **Calculate the Variance:**

* Find the average of these squared deviations.
* For a **population variance** (if you're considering the data as the entire population):



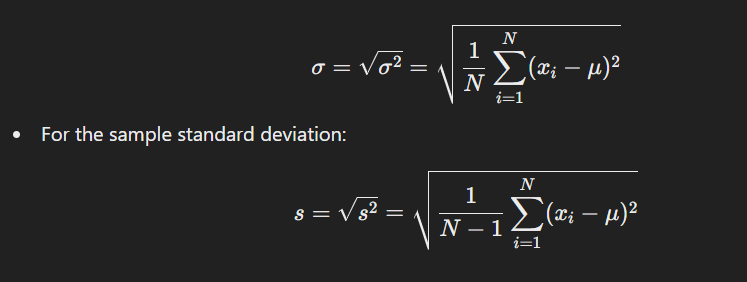
* For a **sample variance** (if you're considering the data as a sample of a larger population):



The sample variance uses N−1N-1N−1 in the denominator (known as Bessel's correction) to correct for the bias that occurs when estimating a population parameter from a sample.

### 5. **Calculate the Standard Deviation:**

* Take the square root of the variance to get the standard deviation.
* For the population standard deviation:



### Summary

* **Variance** gives a measure of how spread out the data points are around the mean.
* **Standard deviation** is the square root of the variance and provides a measure of the average distance of each data point from the mean, in the same units as the data.

These calculations help in understanding the dispersion or variability within a dataset.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**What is skewness and kurtosis? How do they help in understanding data distribution?**

### Skewness

**Skewness** is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. In simpler terms, it tells us whether the data is skewed to the left (negatively skewed), to the right (positively skewed), or if it's symmetrical (zero skewness).

* **Positive Skewness (Right-Skewed):** If the tail on the right side of the distribution is longer or fatter than the left side, the distribution has positive skewness. This indicates that there are more lower values and a few very high values pulling the mean to the right.
* **Negative Skewness (Left-Skewed):** If the tail on the left side of the distribution is longer or fatter than the right side, the distribution has negative skewness. This indicates that there are more higher values and a few very low values pulling the mean to the left.
* **Zero Skewness (Symmetrical Distribution):** If the distribution is perfectly symmetrical, it has zero skewness. This is typical of a normal distribution.

### Kurtosis

**Kurtosis** is a measure of the "tailedness" of the probability distribution of a real-valued random variable. It describes how heavy or light the tails of the distribution are compared to a normal distribution.

* **Leptokurtic (Positive Kurtosis):** Distributions with positive kurtosis have heavier tails than a normal distribution. This means there are more extreme values (outliers) than in a normal distribution.
* **Platykurtic (Negative Kurtosis):** Distributions with negative kurtosis have lighter tails than a normal distribution. This means there are fewer extreme values, and the distribution is flatter and more spread out.
* **Mesokurtic (Zero Kurtosis):** A mesokurtic distribution has a kurtosis value of zero, indicating that it has the same tailedness as a normal distribution.

### How They Help in Understanding Data Distribution

* **Skewness** provides insight into the symmetry of the data distribution. If data is skewed, it suggests that the mean is being influenced by outliers in one direction, which can affect statistical analyses like the mean, median, and mode.
* **Kurtosis** gives information about the presence of outliers and the shape of the tails. High kurtosis implies a higher likelihood of outliers, which can impact the reliability of certain statistical tests and models. Conversely, low kurtosis indicates a more uniform ditribution without many extreme deviations.

Together, skewness and kurtosis help in understanding the shape and characteristics of the data distribution, which is crucial for choosing appropriate statistical methods and interpreting results accurately.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**What is the central limit theorem, and why is it important in statistics?**

The **Central Limit Theorem (CLT)** is a fundamental theorem in statistics that describes how the distribution of sample means approximates a normal distribution, regardless of the original population's distribution, as the sample size becomes large.

### **Key Points of the Central Limit Theorem:**

1. **Sample Means Distribution:** If you take a sufficiently large number of samples from a population with any shape of distribution (normal or not) and calculate their means, the distribution of those means will approximate a normal distribution (a bell curve).
2. **Population Mean and Variance:** The mean of the sample means will be equal to the population mean. The variance of the sample means will be equal to the population variance divided by the sample size (σ^2/n).
3. **Sample Size:** The larger the sample size, the closer the distribution of the sample means will be to a normal distribution. Typically, a sample size of 30 or more is considered sufficient for the CLT to hold.

### **Why is the Central Limit Theorem Important?**

1. **Justification for Using Normal Distribution:** Many statistical methods, such as confidence intervals and hypothesis testing, rely on the assumption that the data are normally distributed. The CLT provides the justification for this assumption when working with sample means, even if the original data does not follow a normal distribution.
2. **Simplifies Analysis:** Since the distribution of sample means approaches normality as sample size increases, it allows for the use of normal probability tables and other tools that assume normality. This simplifies analysis and makes it more accessible.
3. **Foundation for Inferential Statistics:** The CLT is the backbone of inferential statistics. It allows statisticians to make inferences about population parameters (like the mean) based on sample data, which is crucial for making predictions, decisions, and generalizations from a limited amount of data.

In summary, the Central Limit Theorem is crucial because it enables statisticians to use the normal distribution as a model for the distribution of sample means, facilitating the application of various statistical methods and making it easier to draw reliable conclusions from data.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## **Describe different types of probability distributions (e.g., normal, binomial, Poisson).**

Probability distributions describe how the values of a random variable are distributed. Each type of probability distribution has its own characteristics, which make it suitable for modeling different types of random processes. Here's a look at some common probability distributions:

### 1. **Normal Distribution**

* **Concept:** The normal distribution, also known as the Gaussian distribution, is a continuous probability distribution characterized by its bell-shaped curve. It is symmetric around the mean, and most of the data points cluster around the central peak, with fewer points appearing as you move away from the mean.
* **Key Features:**
  + Mean (µ) determines the center of the distribution.
  + Standard deviation (σ) determines the spread or width of the distribution.
  + Many natural phenomena (like heights, test scores, etc.) follow a normal distribution.

### 2. **Binomial Distribution**

* **Concept:** 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 only two possible outcomes (success or failure). It is often used to model situations like the number of heads in a series of coin flips.
* **Key Features:**
  + The probability of success (p) and failure (1-p) is constant for each trial.
  + The number of trials (n) is fixed.
  + The distribution is characterized by the parameters n and p.

### 3. **Poisson Distribution**

* **Concept:** The Poisson distribution is a discrete probability distribution that models the number of times an event occurs in a fixed interval of time or space, given that these events happen with a known constant mean rate and independently of the time since the last event. It is often used to model rare events, such as the number of emails received in an hour or the number of accidents at an intersection.
* **Key Features:**
  + Defined by the average rate (λ) at which events occur.
  + The probability of more than one event occurring in an infinitesimally small time interval is negligible.
  + The events are independent.

### 4. **Uniform Distribution**

* **Concept:** The uniform distribution can be either discrete or continuous. In its continuous form, it models a situation where all outcomes in a range are equally likely. For example, the probability of getting any number between 0 and 1 in a random draw is the same.
* **Key Features:**
  + For a continuous uniform distribution between a and b, any value within [a, b] is equally likely.
  + The probability density function (PDF) is constant.

### 5. **Exponential Distribution**

* **Concept:** The exponential distribution is a continuous probability distribution often used to model the time between independent events that happen at a constant average rate. It's closely related to the Poisson distribution and is commonly used in scenarios like the time until the next customer arrives or the time until a light bulb burns out.
* **Key Features:**
  + Characterized by the rate parameter (λ), which is the inverse of the mean.
  + The distribution is memoryless, meaning the probability of an event occurring in the future is independent of the past.

### 6. **Chi-Square Distribution**

* **Concept:** The chi-square distribution is a continuous distribution that arises in the context of hypothesis testing and confidence interval estimation for variance. It is the distribution of the sum of the squares of independent standard normal variables.
* **Key Features:**
  + Used in statistical tests like the chi-square test for independence.
  + The shape of the distribution depends on the degrees of freedom.

### 7. **Student's t-Distribution**

* **Concept:** The t-distribution is a continuous probability distribution that is similar to the normal distribution but has heavier tails. It is used when the sample size is small and the population standard deviation is unknown, which makes it useful in estimating population parameters.
* **Key Features:**
  + It has a mean of 0 and is symmetric.
  + The spread of the distribution depends on the degrees of freedom.
  + As the sample size increases, the t-distribution approaches the normal distribution.

### 8. **Gamma Distribution**

* **Concept:** The gamma distribution is a continuous probability distribution used to model the time until an event happens a certain number of times. It generalizes the exponential distribution by allowing the shape parameter to take on non-integer values.
* **Key Features:**
  + Defined by a shape parameter (k) and a rate parameter (θ).
  + Useful in various fields such as queuing models and reliability analysis.

### 9. **Beta Distribution**

* **Concept:** The beta distribution is a continuous probability distribution that models the distribution of probabilities themselves. It's often used to model the behavior of random variables that are constrained within an interval, like proportions or percentages.
* **Key Features:**
  + Defined by two shape parameters, α and β.
  + The distribution can take on various shapes depending on the values of α and β, including uniform, U-shaped, or bell-shaped.

Each of these distributions is suited for modeling different types of real-world phenomena, and choosing the appropriate distribution depends on the nature of the data and the problem at hand.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## **Explain the difference between a population and a sample. Why is sampling important?**

### Difference Between a Population and a Sample

* **Population**: A population includes all members or items that meet a particular criterion or set of criteria. It is the entire group about which you want to draw conclusions. For example, if you're studying the average height of adult men in a country, the population would include every adult man in that country.
* **Sample**: A sample is a subset of the population that is selected for the actual study or analysis. The sample should ideally represent the population from which it is drawn. For instance, if you can't measure the height of every adult man in the country, you might select a few thousand men as a sample to estimate the average height.

### Importance of Sampling

1. **Feasibility**: In many cases, it is impractical or impossible to study an entire population due to time, cost, or logistical constraints. Sampling allows researchers to gather and analyze data more efficiently.
2. **Cost-Effective**: Studying a smaller group (sample) instead of an entire population reduces the resources needed, making research more affordable.
3. **Time-Efficient**: Analyzing a sample is quicker than analyzing an entire population, which speeds up the research process and allows for faster decision-making.
4. **Manageability**: Working with a smaller sample makes data collection, storage, and analysis more manageable. Large data sets can be complex and challenging to work with.
5. **Accuracy**: If done correctly, sampling can provide accurate estimates of population parameters. Proper sampling techniques ensure that the sample represents the population, allowing for generalization of the results.
6. **Ethical Considerations**: In some cases, it may be unethical to study an entire population (e.g., testing a new drug on all patients). Sampling allows researchers to conduct studies in a way that minimizes harm or inconvenience.

In summary, while a population encompasses all members of a defined group, a sample is a manageable subset used for study. Sampling is crucial because it makes research feasible, cost-effective, and accurate while allowing researchers to draw meaningful conclusions about a larger group.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## **How to Formulate Null and Alternative Hypotheses?**

1. **Identify the Research Question:**
   * Start with a clear research question or problem you want to investigate.
   * Determine what you want to prove or test.
2. **Formulate the Null Hypothesis (H₀):**
   * Assume no effect, no difference, or no change. This hypothesis should be stated in a way that can be tested directly.
   * Example: "The average weight loss from the new diet is equal to the average weight loss from the existing diet."
   * **H₀:** The population parameter equals a specific value (e.g., H0:μ=0 if testing for no difference).
3. **Formulate the Alternative Hypothesis (H₁ or Hₐ):**
   * This hypothesis should reflect the claim or effect you want to investigate. It is usually the opposite of the null hypothesis.
   * Example: "The average weight loss from the new diet is different from the average weight loss from the existing diet."
   * **H₁:** The population parameter differs from the value stated in H0​ (e.g., H1:μ≠0 for a two-tailed test).
4. **Determine the Type of Test (One-tailed or Two-tailed):**
   * **One-tailed test**: Used when you are testing for an effect in a specific direction (e.g., greater than or less than).
   * **Two-tailed test**: Used when you are testing for an effect in either direction (e.g., not equal).
5. **Set Up the Hypotheses:**
   * **Example:** Suppose a company claims that their light bulbs last for 1000 hours on average. You want to test this claim.
     + **Null Hypothesis (H₀):** The average life of the light bulbs is 1000 hours ( H0:μ=1000 ).
     + **Alternative Hypothesis (H₁):** The average life of the light bulbs is not 1000 hours ( H1:μ≠1000 ).

In summary, the null hypothesis is the claim you seek to test (often a statement of no effect), and the alternative hypothesis is what you are trying to find evidence for (a statement of an effect or difference).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## **Describe the steps in conducting a hypothesis test.**

Conducting a hypothesis test involves several key steps, which are designed to assess the evidence against a null hypothesis. Here’s a step-by-step guide to performing a hypothesis test:

### 1. **State the Hypotheses**

* **Null Hypothesis (H₀):** This is the default assumption that there is no effect or no difference. It represents the status quo or the claim to be tested.
* **Alternative Hypothesis (H₁ or Hₐ):** This is what you want to prove. It represents a new effect, difference, or relationship that contradicts the null hypothesis.

Example:

* H₀: The mean of the population is equal to 50.
* H₁: The mean of the population is not equal to 50.

### 2. **Choose the Significance Level (α)**

* The significance level, denoted as α, is the probability of rejecting the null hypothesis when it is actually true. Common choices for α are 0.05, 0.01, or 0.10, with 0.05 being the most common.

Example:

* Set α = 0.05.

### 3. **Select the Appropriate Test**

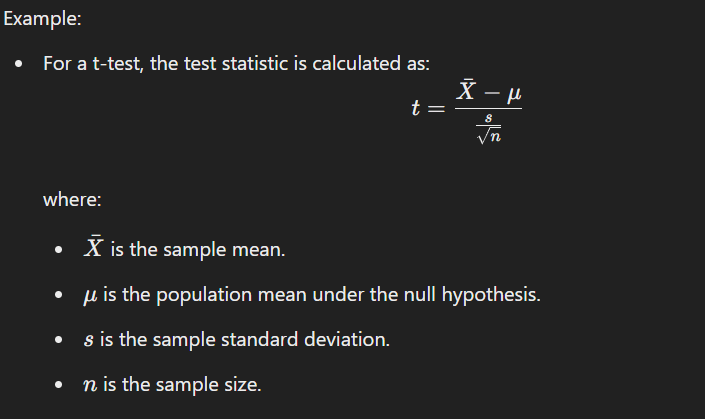
* Choose the statistical test that matches your data and hypothesis. The choice depends on factors like the type of data (e.g., categorical or continuous), sample size, and whether you are comparing means, proportions, etc.

Common tests include:

* **Z-test or t-test:** For comparing means.
* **Chi-square test:** For categorical data.
* **ANOVA:** For comparing means across multiple groups.
* **Regression analysis:** For relationships between variables.

### 4. **Calculate the Test Statistic**

* The test statistic is a standardized value that measures the degree of departure from the null hypothesis. The formula and method for calculating the test statistic depend on the chosen test.



### 5. **Determine the p-value or Critical Value**

* **p-value:** The p-value is the probability of obtaining a test statistic at least as extreme as the one observed, assuming the null hypothesis is true. A small p-value indicates strong evidence against H₀.
* **Critical Value:** Alternatively, you can compare the test statistic to a critical value from a statistical distribution (e.g., t-distribution, Z-distribution) based on the chosen significance level.

Example:

* If using a p-value approach, calculate the p-value associated with the test statistic.
* If using a critical value approach, determine the critical value from statistical tables and compare it to the test statistic.

### 6. **Make a Decision**

* **Reject H₀:** If the p-value is less than or equal to α, or if the test statistic exceeds the critical value, reject the null hypothesis. This suggests there is sufficient evidence to support the alternative hypothesis.
* **Fail to Reject H₀:** If the p-value is greater than α, or if the test statistic does not exceed the critical value, fail to reject the null hypothesis. This suggests there is not enough evidence to support the alternative hypothesis.

### 7. **Draw a Conclusion**

* Based on the decision, conclude whether or not there is sufficient statistical evidence to support the alternative hypothesis. The conclusion should be stated in the context of the original research question or problem.

### 8. **Report the Results**

* Clearly report the test statistic, p-value, decision (whether H₀ was rejected), and the conclusion. It’s also helpful to include the confidence interval for the estimated effect size, if applicable.

Example:

* "The t-test results show that the mean is significantly different from 50 (t = 2.45, p = 0.017). Thus, we reject the null hypothesis at the 0.05 significance level.

Following these steps ensures a systematic and rigorous approach to hypothesis testing, allowing for sound conclusions based on statistical evidence.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
9. What is a p-value? How do you interpret it in the context of a hypothesis test? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
10. When would you use a t-test versus a z-test?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
11. Explain how you would conduct an independent two-sample t-test. What assumptions must be met?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
12. Describe a scenario where you would use a paired sample t-test.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
13. What is ANOVA, and how does it differ from a t-test?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
14. Explain how you would interpret the results of a one-way ANOVA.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
15. Describe a situation where you might use a two-way ANOVA.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
16. What is a chi-square test for independence? When would you use it?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
17. How do you interpret the results of a chi-square goodness-of-fit test?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
18. Explain the assumptions and limitations of chi-square tests.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
19. What is the difference between simple linear regression and multiple regression?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
20. How do you assess the goodness-of-fit of a regression model?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
21. Explain multicollinearity and how you would detect and handle it in a regression model.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
22. What is the difference between correlation and causation?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
23. How do you interpret the Pearson correlation coefficient?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
24. When would you use Spearman rank correlation instead of Pearson correlation?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
25. What are some common methods for forecasting time series data?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
26. Explain the components of a time series (trend, seasonality, residuals).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
27. How would you handle missing data in a time series dataset?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
28. Describe your approach to exploratory data analysis (EDA).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

29. How do you handle outliers in a dataset?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
30. Explain the steps you would take to validate the results of your analysis.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
31. Give an example of how you have used statistical analysis to solve a real-world problem

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_