**Capstone Project: Applied Statistics**

1. What is a vector in mathematics?

Ans: In mathematics, a vector is a quantity that has both **magnitude** (size/length) and **direction**. It is often represented as an ordered set of numbers or coordinates in a coordinate system. Vectors are fundamental in various fields such as physics, engineering, and computer science, and they are used to represent quantities like force, velocity, and displacement.

Think of it as an arrow pointing in a specific direction with a certain length.

Magnitude: This is the length of the arrow, representing the strength or intensity of the quantity.

Direction: The arrow points in a specific direction, indicating which way the quantity is acting.

1. How is a vector different from a scalar?

Ans: A vector is different from a scalar in that a scalar only has magnitude, while a vector has both magnitude and direction.

Common examples of scalars include quantities like temperature, mass, time, and distance. For instance, 5 kilograms of weight or 20 degrees Celsius of temperature.

Examples of vectors include quantities like velocity, force, and displacement. For instance, a velocity vector might be represented as 𝑣=(3,4) in two-dimensional space

1. What are the different operations that can be performed on vectors?

Ans: i) Adding two vectors results in a new vector is called as vector addition

Component-wise addition. If 𝑢 = (𝑢1, 𝑢2) and 𝑣 = (𝑣1, 𝑣2) then 𝑢+𝑣=(𝑢1+𝑣1,𝑢2+𝑣2)

ii) Subtracting one vector from another results in a new vector is called as vector subtraction. Component-wise subtraction. If 𝑢=(𝑢1,𝑢2)and 𝑣=(𝑣1,𝑣2), then 𝑢−𝑣=(𝑢1−𝑣1,𝑢2−𝑣2)**.**

iii) Multiplying a vector by a scalar (a real number) is called as scalar multiplication.

Each component of the vector is multiplied by the scalar. If 𝑐 is a scalar and 𝑣=(𝑣1,𝑣2) then 𝑐𝑣=(𝑐𝑣1,𝑐𝑣2)

### iv) An operation that takes two vectors and returns a scalar is a Dot Product (Scalar Product). Multiply corresponding components and sum the results. For 𝑢=(𝑢1,𝑢2,𝑢3) and 𝑣=(𝑣1,𝑣2,𝑣3) the dot product is 𝑢⋅𝑣=𝑢1𝑣1+𝑢2𝑣2+𝑢3𝑣3

v) An operation in three-dimensional space that takes two vectors and returns a new vector perpendicular to both is cross product.

For 𝑢 = (𝑢1, 𝑢2, 𝑢3) and 𝑣 = (𝑣1, 𝑣2, 𝑣3), the cross product is:

𝑢×𝑣 = (𝑢2𝑣3−𝑢3𝑣2, 𝑢3𝑣1−𝑢1𝑣3, 𝑢1𝑣2−𝑢2𝑣1)

vi) Projecting one vector onto another is vector projection.

1. How can vectors be multiplied by a scalar?

Ans: When a vector is multiplied by a scalar, each component of the vector is multiplied by the scalar value.

For example, if we have a vector A = (a1, a2, a3) and a scalar c, the scalar multiplication of the vector is c \* A = (c \* a1, c \* a2, c \* a3).

1. What is the magnitude of a vector?

Ans: The magnitude of a vector, often referred to as its length or size, is a measure of how long the vector is. It is a scalar value representing the distance of the vector from the origin to its endpoint in a coordinate system.

The magnitude of a vector can be calculated using the Pythagorean theorem

A = (a1, a2, a3) is given by ||A|| = √(a1^2 + a2^2 + a3^2).

1. How can the direction of a vector be determined?

Ans: **Using Trigonometric Functions (for 2D vectors):** If you know the vector's components (i.e., its changes in X and Y coordinates), you can use the tangent function (tan) to find the angle (theta, θ) representing its direction.

### Using Unit Vectors

A unit vector points in the same direction as the original vector but has a magnitude of 1. To find the unit vector 𝑣^**,** you divide the vector by its magnitude.

1. What is the difference between a square matrix and a rectangular matrix?

Ans: **Square matrix:** This type of matrix has the same number of rows and columns. For instance, a matrix with dimensions (3x3) or (5x5) would be considered square matrices.

**Rectangular matrix:** In contrast, a rectangular matrix has a different number of rows compared to columns. Examples of rectangular matrices include (2x3) and (4x7) matrices.

Square matrices have special properties (like determinants and eigenvalues) that rectangular matrices do not inherently possess.

1. What is a basis in linear algebra?

Ans: A basis is a set of linearly independent vectors that can be used to express any other vector in a given vector space.

The vectors in the set are **linearly independent**, meaning that no vector in the set can be written as a linear combination of the other vectors.

**Spanning the Space:** Every single vector within the vector space can be expressed as a linear combination (a weighted sum) of the vectors in the basis.

By having a basis, you can represent any vector in the space using coordinates (the weights used in the linear combination). This allows you to analyze and manipulate vectors more efficiently.

The number of vectors in a basis for a particular vector space is always the same, and this number is called the dimension of the space. So, a basis essentially tells you how many fundamental directions or "degrees of freedom" there are within the space.

1. What is a linear transformation in linear algebra?

Ans: A linear transformation is a function that maps one vector space to another in a way that preserves the basic structure of the space.

a function between two vector spaces that preserves the operations of vector addition and scalar multiplication.

They are often represented by matrices. Multiplying a vector by a matrix is a common way to perform a linear transformation.

Linear transformations are fundamental in linear algebra because they allow us to study how different structures like vectors and matrices relate to each other in a predictable way. They are essential for many areas of mathematics, physics, engineering, and computer science and machine learning.

10. What is an eigenvector in linear algebra?

Ans: An eigenvector is a non-zero vector that, when multiplied by a given square matrix, results in a scalar multiple of the original vector. The scalar multiple is called the eigenvalue.

In linear algebra, an eigenvector of a square matrix 𝐴 is a non-zero vector 𝑣 such that when 𝐴 is multiplied by 𝑣, the result is a scalar multiple of 𝑣*.* Formally, 𝑣 is an eigenvector of 𝐴 if there exists a scalar 𝜆 such that:

𝐴𝑣=𝜆𝑣

The scalar 𝜆is called the eigenvalue corresponding to the eigenvector 𝑣.

11. What is probability theory?

Ans: Probability theory is a branch of mathematics concerned with analyzing random phenomena and events. The main goal is to quantify the likelihood of various outcomes and understand the patterns and behaviors of systems involving uncertainty.

**Random events:** These are events where the outcome can't be predicted for certain. Flipping a coin, rolling a die, or drawing a card from a deck are all examples of random events.

**Sample space:** This refers to the set of all possible outcomes of an experiment. For example, the sample space for flipping a coin is {heads, tails}.

**Probability:** This is a numerical measure between 0 and 1 that represents the likelihood of an event happening. 0 indicates the event is impossible, and 1 signifies certainty.

12. What are the primary components of probability theory?

Ans: The primary components of probability theory include the foundational concepts and mathematical structures that define and quantify randomness and uncertainty.

The primary components of probability theory include probability axioms and rules, conditional probability and Bayes theorem, random variables, and the law of large numbers and central limit theorem.

13. What is conditional probability, and how is it calculated?

Ans: Conditional probability refers to the likelihood of one event (let's call it B) happening, given that another event (A) has already occurred. It's basically updating the probability of event B after you know something extra (that event A happened).

P(B|A) represents the probability of event B occurring, **given that** event A already happened.

The numerator, P(A∩B), represents the probability of both event A and B happening together. This is called the intersection of events A and B.

The denominator, P(A), represents the probability of event A happening in the first place.

Here's the formula for conditional probability:

P(B|A) = P(A∩B) / P(A)

#### 14. What is Bayes theorem, and how is it used?

Ans: Bayes theorem is a mathematical formula that calculates the probability of an event based on prior knowledge or observations. It updates our beliefs about an event based on new information.

**Concept:** P(A|B) = (P(B|A) \* P(A)) / P(B)

P(A|B) represents the probability of event A happening given event B already occurred (what we want to find).

P(B|A) represents the probability of event B happening, given that event A already happened (often easier to determine).

P(A) represents the initial probability of event A happening before considering event B (also known as the prior probability).

P(B) represents the probability of event B happening in general (regardless of A).

In simpler terms, Bayes' theorem allows you to revise the initial probability (prior probability) of event A by taking into account the new information provided by event B.

15. What is a random variable, and how is it different from a regular variable?

Ans: A random variable is a fundamental concept in probability and statistics that represents a numerical outcome of a random phenomenon. Unlike regular variables, which typically take on specific values deterministically, random variables take on values according to a probability distribution.

Regular variables and random variables are both used to represent unknowns, but they differ in how their values are determined:

For example, in the equation x + 5 = 10, x is a regular variable. It can take on any value that satisfies the equation, in this case, x = 5,

if you flip a coin, the random variable could be "heads" or "tails," each with a probability of 0.5.

16. What is the law of large numbers, and how does it relate to probability theory?

Ans: The law of large numbers states that as the sample size of a random variable increases, the sample mean converges to the population mean. It is a fundamental concept in probability theory used to make predictions based on large datasets.

Essentially, as the number of trials in a random experiment increases, the average of the results gets increasingly closer to the expected value (or theoretical mean).

**Probability theory provides the framework for understanding expected values and the likelihood of different outcomes in random events.** The law of large numbers builds upon this foundation by demonstrating a **predictable pattern** that emerges in random experiments with a large number of trials.

#### 17. What is the central limit theorem, and how is it used?

Ans : The CLT states that under certain conditions, as the sample size gets larger, the distribution of the sample means approaches a normal distribution (bell curve), regardless of the original population's distribution shape.

Imagine you have a population with heights that follow a non-normal distribution, maybe skewed towards shorter people. If you take small samples from this population (say, 5 people), the average height of each sample might also not be normally distributed. But the CLT tells us that if you keep increasing the sample size (say, 30 or more people), the distribution of these sample means will start to resemble a normal distribution, even though the original population wasn't normal!

18. . What is the difference between discrete and continuous probability distributions?

Ans :

**Discrete Probability Distribution:**

A random variable is considered discrete if it can only take on a finite or countable number of distinct values. These values are usually integers.

Examples: Number of times a die rolls a 6, number of customers arriving in a store in an hour, number of correct answers on a multiple-choice exam.

Discrete probability distributions are often represented using probability mass functions (PMFs). A PMF assigns a probability to each possible value of the random variable.

**Continuous Probability Distribution:**

A random variable is considered continuous if it can take on any value within a specific range (which can be finite or infinite). These values are typically measured along a continuous scale.

Examples: Height of a person, time it takes to complete a marathon, weight of a car.

Continuous probability distributions are described using probability density functions (PDFs). A PDF describes the probability density over a range of values, rather than assigning a specific probability to each individual value.

19. What are some common measures of central tendency, and how are they calculated?

Ans : Measures of central tendency are statistical metrics that describe the center or typical value of a dataset. The three most common measures of central tendency are the mean, median, and mode.

**Mean:** The most common measure, often referred to as the "average," is the sum of all the values in the data set divided by the number of values.

**Median:** The median represents the middle value when the data is ordered from least to greatest. If you have an even number of data points, the median is the average of the two middle values.

**Mode:** The mode is the most frequent value in the data set. A data set can have multiple modes (bimodal) or no mode at all.

20. What is the purpose of using percentiles and quartiles in data summarization?

Ans: Percentiles and quartiles are valuable tools in data summarization because they offer a more comprehensive picture of how data is distributed compared to just using the mean or median

Percentiles and quartiles help to identify the spread of values within a dataset, particularly when dealing with skewed distributions. They can help identify values that fall above or below a certain threshold and provide insight into the distribution of values.

**Percentiles:**

Divide your data set into 100 equal parts. Each percentile represents the value below which a certain percentage of data points fall.

Useful for comparing data points to specific percentiles. Knowing someone scored in the 95th percentile on an exam tells you they did better than 95% of test-takers.

**Quartiles:**

Divide your data set into four equal parts. The quartiles (Q1, Q2, Q3) represent the values that separate the data into these four quarters.

Summarize how your data is spread out by providing key reference points that divide the data into quarters.

21. How do you detect and treat outliers in a dataset?

Ans: Outliers are data points that fall significantly outside the overall pattern of the rest of the data. They can arise due to errors in measurement, recording, or natural variations in the data.

**Data Visualization:**

Scatter plots and box plots are excellent tools for visually identifying outliers. In scatter plots, outliers will appear as distant points from the main cluster of data points. Box plots show the interquartile range (IQR) and can easily reveal data points falling outside the whiskers (1.5 IQR from Q1 and Q3).

**Statistical Methods:**

Z-scores: These compare each data point to the mean and standard deviation. Values more than 3 standard deviations away from the mean are considered potential outliers.

Interquartile Range (IQR): As mentioned earlier, IQR is the difference between Q3 (75th percentile) and Q1 (25th percentile). Data points below Q1 - 1.5 IQR or above Q3 + 1.5 IQR are potential outliers.

22. How do you test the goodness of fit of a discrete probability distribution?

Ans: Testing the goodness of fit of a discrete probability distribution involves assessing how well the observed data fits a theoretical or expected distribution. Several statistical tests can be used for this purpose, depending on the nature of the data and the distribution being tested.

The chi-square (𝜒2) test is a widely used method for testing the goodness of fit of a discrete probability distribution. It compares observed frequencies with expected frequencies under the null hypothesis that the observed data follow the specified distribution.

23. What is a joint probability distribution?

Ans: A joint probability distribution is a probability distribution that describes the simultaneous occurrence of two or more random variables. It specifies the probability of all possible outcomes of the random variables together.

Joint probability distributions are crucial in various fields, including statistics, machine learning, and decision theory. They provide a comprehensive understanding of the relationship between multiple random variables and enable the calculation of various probabilities and statistical measures, such as marginal probabilities, conditional probabilities, and expectations.

#### 24. What is the difference between a joint probability distribution and a marginal probability distribution?

Ans: A joint probability distribution describes the probabilities of two or more random variables occurring simultaneously. In contrast, a marginal probability distribution describes the probabilities of a single random variable occurring regardless of the other variables. Think of a joint probability distribution as a detailed weather forecast table showing the probability of rain/shine and temperature (low/high) for each day of the week. The marginal probability distribution for rain/shine would be like looking at just the rain/shine probabilities for the entire week, ignoring the temperature information.

#### 25. What is the covariance of a joint probability distribution?

Ans : The covariance of a joint probability distribution measures the degree to which two random variables are related.

It is calculated using the formula

Cov(X, Y) = E(XY) - E(X)E(Y),

where E(XY) is the expected value of the product of X and Y, and

E(X) and E(Y) are the expected values of X and Y, respectively.

This formula provides a measure of how the two variables co-vary, or vary together. If the covariance is positive, it indicates that the variables tend to move in the same direction. If it's negative, they tend to move in opposite directions. If it's zero, there's no linear relationship between them.

#### 26. What is the relationship between the correlation coefficient and the covariance of a joint probability distribution?

Ans: The correlation coefficient is a standardized version of the covariance, which allows for comparison of the degree of relationship between variables on different scales.

By dividing the covariance by the product of the standard deviations, the correlation coefficient is scaled to be dimensionless and ranges between -1 and +1, making it easier to interpret and compare across different datasets.

It is calculated using the formula

ρ(X, Y) = Cov(X, Y) / (σ(X)σ(Y)),

where σ(X) and σ(Y) are the standard deviations of X and Y, respectively.

If 𝜌(𝑋,𝑌)=1it indicates a perfect positive linear relationship between 𝑋 and 𝑌.

If 𝜌(𝑋,𝑌)=−1, it indicates a perfect negative linear relationship between 𝑋 and 𝑌.

If 𝜌(𝑋,𝑌)=0, it indicates no linear relationship between 𝑋 and 𝑌.

In summary, while covariance measures the degree to which two variables vary together, the correlation coefficient standardizes this measure, making it easier to interpret and compare across different datasets.

27. How do you determine if two random variables are independent based on their joint probability distribution?

Ans: Two random variables 𝑋and 𝑌are considered independent if their joint probability distribution can be factored into the product of their marginal probability distributions.

covariance (cov(X, Y)) measures the linear relationship between two variables. If the covariance is zero, it suggests no linear relationship, which can be a sign of independence (but not a definitive confirmation). However, keep in mind that:

Covariance only detects linear relationships. Even if the covariance is zero, the variables might still be dependent in a non-linear way.

28. What is sampling in statistics, and why is it important?

Ans: Sampling refers to the process of selecting a subset of individuals or items from a larger population.

It is important because it allows us to make inferences about the entire population based on the characteristics observed in the sample.

Sampling helps in reducing the time, cost, and effort required to collect data from the entire population.

By understanding sampling techniques and their importance, you can leverage them to draw reliable conclusions about populations from a manageable amount of data. This is fundamental for various fields like public opinion research, marketing studies, scientific experiments, and quality control.

29. What are the different sampling methods commonly used in statistical inference?

Ans: i) **Simple Random Sampling**: Every member of the population has an equal chance of being selected

ii) **Stratified Sampling**: The population is divided into distinct subgroups or strata, and samples are then randomly selected from each stratum. This ensures representation from each subgroup.

ii) **Systematic Sampling**: Sampling units are selected at regular intervals from an ordered list of the population. For example, selecting every 10th person from a voter list.

iv) **Cluster Sampling**: The population is divided into clusters, usually based on geographical location, and then a random sample of clusters is selected. All members within the chosen clusters are included in the sample.

v) **Convenience Sampling**: This involves selecting the most readily available members of the population. While convenient, this method may introduce bias as it may not be representative of the entire population.

#### 30. What is the central limit theorem, and why is it important in statistical inference?

Ans: The central limit theorem states that when independent random variables are added, their sum tends toward a normal distribution, regardless of the shape of the original distribution.

The importance of the Central Limit Theorem in statistical inference lies in its implications for making inferences about population parameters based on sample statistics.

**Allows for Inference about Population Parameters**: The CLT enables statisticians to make inferences about population parameters, such as the population mean or proportion, based on sample data. By knowing that the sampling distribution of the sample mean will be approximately normal, even if the population distribution is not, we can use properties of the normal distribution to make inferences.

**Foundation for Hypothesis Testing and Confidence Intervals**: Many statistical methods, including hypothesis testing and confidence interval estimation, rely on the assumption of normality.

#### 31. What is the difference between parameter estimation and hypothesis testing?

Ans: **Parameter estimation** aims to determine the approximate value of a population parameter (e.g., mean, variance, proportion) based on sample data. Focuses on providing an estimate of a population parameter.

**Hypothesis testing** is used to make decisions about the validity of a specific claim (hypothesis) regarding a population parameter based on sample data. Focuses on assessing the validity of a specific hypothesis about a population parameter.

#### 32. What is the p-value in hypothesis testing?

Ans: The p-value in hypothesis testing is a measure of the strength of evidence against the null hypothesis. The p-value is the probability of obtaining a test statistic as extreme as, or more extreme than, the observed value, assuming that the null hypothesis is true. It is used to determine the statistical significance of the results and helps in deciding whether to reject or fail to reject the null hypothesis.

A smaller p-value indicates stronger evidence against the null hypothesis, **High p-value indicates weaker evidence** leading to a failure to reject the null hypothesis.

33. What is confidence interval estimation?

Ans: Confidence interval estimation is a statistical technique used to estimate a range within which a population parameter (such as a mean, proportion, or variance) is likely to lie with a certain level of confidence.

Unlike point estimates, which provide a single value, confidence intervals provide a range of plausible values for the parameter, giving more information about the uncertainty associated with the estimate.

34. What are Type I and Type II errors in hypothesis testing?

Ans: A Type I error occurs when the null hypothesis is true, but we incorrectly reject it.

his means we conclude that there is an effect or a difference when, in reality, there isn't one. It represents a false positive result. The probability of Type I error is denoted by the significance level (α).

A Type II error occurs when the null hypothesis is false, but we fail to reject it.

This means we conclude that there is no effect or difference when, in reality, there is one. It represents a false negative result. The probability of Type II error is denoted by β.

35. What is the difference between correlation and causation?

Ans: Correlation refers to a statistical relationship between two variables, indicating how they move together. Indicates a statistical association between variables.

Causation, on the other hand, implies that one variable directly influences or causes a change in another variable. Indicates a direct cause-effect relationship between variables.

While correlation can suggest a potential relationship, it does not prove causation, as there may be other underlying factors or confounding variables at play.

36. What does the confidence level represent in a confidence interval?

Ans: The confidence level in a confidence interval represents the degree of certainty that the interval contains the true population parameter.

It is a probability, expressed as a percentage, that quantifies how confident we are that the population parameter lies within the calculated interval based on the sample data.

A 95% confidence level means that if we repeated the sampling process 100 times and constructed a confidence interval each time, we would expect approximately 95 of those intervals to contain the true population parameter.

**Not a Probability Statement about the Specific Interval**

The choice of confidence level depends on the desired degree of confidence and the trade-off with interval width. Higher confidence levels (e.g., 99%) yield wider intervals, indicating more uncertainty about the precise location of the parameter but greater assurance that the interval captures the parameter. Lower confidence levels (e.g., 90%) yield narrower intervals but with less assurance.

37. What is hypothesis testing in statistics?

Ans: Hypothesis testing is a statistical method used to make inferences about population parameters based on sample data. It involves formulating a null hypothesis and an alternative hypothesis, collecting sample data, and evaluating the evidence to determine whether there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis.

38. What is the purpose of a null hypothesis in hypothesis testing?

Ans: The null hypothesis represents the default assumption or claim that there is no significant difference or relationship between variables in the population

The null hypothesis provides a standard against which the observed data are compared.

By setting up a null hypothesis, researchers can make informed decisions about the validity of their findings. If the data provide sufficient evidence to reject the null hypothesis, it suggests that there may be a genuine effect or difference.

The null hypothesis helps control the rate of Type I errors (false positives) by setting a threshold (significance level, 𝛼) for rejecting null hypothesis

39. What is the difference between a one-tailed and a two-tailed test?

Ans: The key difference between a one-tailed test and a two-tailed test lies in the **direction** of the effect you're looking for in your hypothesis testing.

Two-tailed test: This is the more general approach. It's used when you expect a deviation from the null hypothesis (H0) in either direction, greater than or less than a certain value.

The alternative hypothesis (H1) is typically stated as "not equal to" (≠) the null hypothesis value.This is suitable when you're unsure of the direction of the effect, or you're open to finding a difference in either direction (positive or negative).

**One-tailed test:** This is used when you have a **specific prediction** about the direction of the effect. You expect the value to be either higher than (> ) or lower than (<) a certain value compared to the null hypothesis.

40. What is experiment design, and why is it important?

**Ans:** Experimental design in statistics refers to the process of planning and conducting experiments to answer research questions or test hypotheses.

Experiment design is the blueprint for conducting a scientific experiment. It outlines the entire process, from formulating a research question to interpreting the results.

Experiment design is important because Ensures reliable and valid results, **Increases clarity and interpretability, Increases clarity and interpretability.**

In essence, experiment design is like building a strong foundation for your scientific investigation.

41. What are some strategies to mitigate potential sources of bias in experiment design?

Ans: Mitigating potential sources of bias in experiment design is crucial for ensuring the validity and reliability of research findings.

**Randomization:** Randomly assign participants to experimental and control groups. This helps ensure any pre-existing group differences are balanced out and the observed effects are more likely due to the independent variable.

**Blinding:** If possible, conduct blind experiments where participants and/or researchers don't know which group belongs to which condition. This reduces bias caused by expectations or interactions.

**Power analysis:** Conduct a power analysis to determine the minimum sample size needed to detect an effect of a certain size with a desired level of confidence.

**Standardized procedures:** Develop clear and detailed protocols for data collection to minimize inconsistencies and subjective interpretations by researchers.

42. How can sample size determination affect experiment design?

Ans: Sample size determination plays a crucial role in experiment design as it directly impacts the reliability, validity, and statistical power of the study.

**Large Sample Sizes:** Larger samples tend to provide more precise estimates of population parameters, leading to greater accuracy in the study findings and also increases the statistical power of the study.

**Small Sample Sizes:** Smaller samples may result in wider confidence intervals and less precise estimates, reducing the accuracy of the findings and increasing the risk of Type II errors (false negatives). They may lower statistical power of the study.

43. How are confidence tests and hypothesis tests similar? How are they different?

Ans: Confidence tests and hypothesis tests are both inferential statistics methods used to draw conclusions about populations based on samples.

Similarities:

Both rely on sample data: Neither method gives you the exact population parameter. They use sample statistics (like mean or proportion) to make inferences about the population.

Both involve sampling distributions: They consider the probability distribution of a statistic (e.g., sample mean) based on repeated sampling from the population.

Differences:

Confidence interval: Estimates a range of values where the population parameter is likely to fall with a certain level of confidence (e.g., 95% confidence interval).

Hypothesis test: Makes a decision about a pre-defined claim regarding a population parameter. It focuses on rejecting or failing to reject a null hypothesis (H0) stating no effect.

44. What is the left-skewed distribution and the right-skewed distribution?

Ans: In statistics, a distribution refers to the pattern of how data points are spread out. When a distribution is skewed, it means the data isn't symmetrical around the center.

**Left-Skewed Distribution (Negative Skew):**

* Imagine the data points piled unevenly on a number line, with a longer tail stretching out towards the **left side**.
* Most of the data values are clustered on the **right side** of the distribution, with fewer values as you move towards the left side.

**Right-Skewed Distribution (Positive Skew):**

* Picture the data points like an uneven pile leaning towards the **right side** of the number line.
* Most of the data values are concentrated on the **left side** of the distribution, with fewer values as you move towards the right side.
* The **mean** is **pulled up** (to the right) compared to the **median**.

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

Ans: **Descriptive Statistics:**

* **Focuses on summarizing and describing the characteristics of a dataset.**
* It provides a clear picture of what's going on within the data itself, without making any claims about a larger population.
* Common methods in descriptive statistics include:
  + Measures of central tendency (mean, median, mode) to understand where the center of the data lies.
  + Measures of dispersion (variance, standard deviation) to quantify how spread out the data is.
  + Charts and graphs (histograms, bar charts, pie charts) to visually represent the data distribution.

**Inferential Statistics:**

* **Aims to draw conclusions about a larger population (population inferences) based on a sample of data.**
* It uses sample data to make generalizations, predictions, or test hypotheses about the population. There's always an element of uncertainty because we're not analyzing the entire population.
* Common methods in inferential statistics include:
  + Hypothesis testing (e.g., t-tests, ANOVA) to assess the likelihood of observing a sample result if a certain claim about the population (null hypothesis) is true.
  + Confidence interval estimation to provide a range of values where the population parameter is likely to fall with a certain level of confidence.

46. What is the probability of throwing two fair dice when the sum is 5 and 8?

**Ans:** Probability of getting a sum of 5:

(1, 4), (2, 3), (3, 2), (4, 1) - Four possible outcomes.

Probability of getting a sum of 8:

(2, 6), (3, 5), (4, 4), (5, 3), (6, 2) - Five possible outcomes.

There are 6×6=36 possible outcomes when rolling two fair dice.

Probability (sum of 5) = Number of favorable outcomes / Total number of outcomes Probability (sum of 5) = 4 / 36

Probability (sum of 8) = Number of favorable outcomes / Total number of outcomes Probability (sum of 8) = 5 / 36

Since these are mutually exclusive events (can't happen on the same roll), we don't simply add them together.

Probability (not sum of 5 and not sum of 8) = 1 - ((4/36) + (5/36))

Probability (not sum of 5 and not sum of 8) = 1 - (9/36)

Probability (not sum of 5 and not sum of 8) = 27/36

Probability (sum of 5 or 8) = 1 - Probability (not sum of 5 and not sum of 8)

Probability (sum of 5 or 8) = 1 - (27/36)

Probability (sum of 5 or 8) = 9/36

Therefore, the probability of throwing two fair dice and getting a sum of 5 or 8 is 9/36, which can be simplified to ¼.

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

**Ans:** In statistics, degrees of freedom (often denoted by v or df) represent the number of independent pieces of information used to estimate a statistic. It essentially reflects how many values in your calculation are free to vary.

**Independent Information:** When calculating a statistic (like mean, variance, etc.) from a sample, not all the data points are completely independent. There's some inherent relationship between them.

**Degrees of Freedom:** Degrees of freedom quantify how many values you can freely choose in your calculation without affecting the others, given the overall sample size.

48. **What is a Chi-Square test?**

Ans: The chi-square test (χ² test) is a statistical hypothesis test used to assess whether two categorical variables are independent of each other. It's a common tool used in analyzing contingency tables, which organize data by frequency counts for different categories.

Chi-square tests are sensitive to sample size. Larger samples tend to produce higher chi-square statistics, even for weak associations.

The chi-square test helps you answer questions like:

Is there a relationship between customer age group and preferred product category?

49. What is the ANOVA test?

Ans: ANOVA, which stands for Analysis of Variance, is a statistical test used to compare the means of more than two groups. It helps determine if there are any statistically significant differences between the groups' averages.

ANOVA uses an F-statistic to compare the between-group variance to the within-group variance.

**How ANOVA breaks down variance:**

**Total Variance:** This represents the total variation in the data, considering all the factors that might contribute to the differences in the dependent variable.

**Within-Group Variance:** This reflects the variation in the dependent variable within each group. It considers factors other than the independent variable that might cause individual data points to deviate from the group's mean (e.g., natural variations in soil quality, minor differences in planting techniques).

**Between-Group Variance:** This captures the variation in the means of the different groups. Ideally, if the independent variable has no effect, this variance should be minimal.

50. 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: we can use the concept of complementary probability,

P(seeing supercar) = 0.3

P(not seeing supercar) = 1 – 0.3 = 0.7

Since we're assuming independence between these intervals (i.e., seeing a supercar in one interval doesn't affect the probability of seeing one in another)

Therefore, the probability of NOT seeing a supercar in the entire 60-minute window is

0.7 ^ 3 = 0.343 …….(60 minutes / 20 minutes/interval = 3 intervals)

we can apply the complementary principle again,

proba­bility that you see at least one supercar in the period of an hour = 1 – 0.343 = **0.657**

So, the probability of seeing at least one supercar in a 60-minute period is approximately **65.7%.**