1. What is a vector in mathematics?

A vector is a mathematical quantity with both magnitude and direction. Here, the magnitude refers to the size of the vector. Geometrically, a vector can be viewed as a directed line segment in which its length is the magnitude of the vector and the arrow indicates the direction. The direction of a vector starts from its tail and ends at its head.

References:

[1] https://byjus.com/maths/vectors/
[2]An introduction to vectors - Math InsightMath Insighthttps://mathinsight.org ›vector_introduction

2. How is a vector different from a scalar?

Based on the direction, physical quantities can be classified into two categories — scalar and vector and are different from each other.
Firstly, scalar quantities only have magnitude and no direction, unlike vectors with both magnitude and direction. Hence, any change in scalar quantity is the reflection of a change in magnitude whereas any change in vector quantity can reflect either a change in direction, change in magnitude, or changes in both. A few examples of scalar quantities include length, mass, energy, density, etc. while displacement, velocity, acceleration, weight, and force are a few examples of vectors.
Secondly, scalars are one-dimensional quantities while vectors can be one, two, or three-dimensional. Thirdly, scalar quantities cannot be resolved as they have the same value regardless of direction, as compared to vector quantities that can be resolved in any direction using the sine or cosine of the adjacent angle.
Fourthly, any mathematical operation carried out among two or more scalar quantities will provide a scalar only. However, if a scalar is operated with a vector then the result will be a vector. On the other hand, the result of mathematical operations between two or more vectors may give either scalar or vector. For example, the dot product of two vectors gives only a scalar; while cross product or summation or subtraction between two vectors results in a vector.

Reference:

[1]https://www.difference.minaprem.com/physics/difference-between-scalar-quantity-and-vector-quantity/

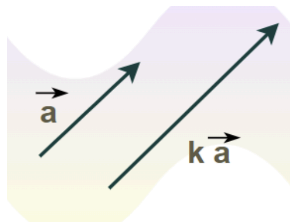
3. What are the different operations that can be performed on vectors?

Various operations can be performed on vector operations such as addition, subtraction, scalar multiplication, dot product, cross product, and finding the magnitude. These operations are fundamental in various mathematical fields and have applications in physics and engineering.

Reference:
[1]https://www.geeksforgeeks.org/vector-operations/

4. How can vectors be multiplied by a scalar?

When the multiplication of a vector is done with a scalar quantity, the magnitude of the vector changes according to the magnitude of the scalar. However, the direction of the vector remains unchanged. For example, if v = [a, b, c] is a vector and k is a scalar, the scalar multiplication would be [ka, kb, kc].



Also, if the scalar quantity multiplied is positive, the direction will be the same. However, for the negative value of the scalar, the direction of the result will be opposite to the original vector. This is also known as the additive inverse of the vector.

References:
[1]https://byjus.com/maths/multiplication-of-vectors-with-scalar/#:~:text=Multiplication%20of%20vectors%20with%20scalar%3A,of%20the%20vector%20remains%20unchanged.
[2] https://www.geeksforgeeks.org/vector-algebra/

5. What is the magnitude of a vector?

The length of a vector is called its magnitude. The magnitude of a vector A is represented using the modulus operator i.e., |A|.
Pythagoras theorem is used to find it out for a vector of any dimension. For example, for a vector defined as xi + yj, the magnitude can be obtained by calculating the square root of the sum of squares of the individual terms. Also, for a vector which starts from the point (x1, y1) and ends at the point (x2, y2), the magnitude is given by the formula:

|A| = √((x2 – x1)2 + (y2 – y1)2)

References:
[1]https://mathinsight.org/definition/magnitude_vector#:~:text=The%20magnitude%20of%20a%20vector,are%20derived%20in%20this%20page.
[2]https://www.geeksforgeeks.org/magnitude-of-a-vector/

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

The orientation of a vector i.e., the angle it makes with the x-axis is termed as its direction. For example, for the velocity vector, its direction denotes where the object is moving. It is represented by the symbol θ and can be calculated by finding the inverse tangent of the ratio of the distance moved by the line along the y-axis to the distance moved along the x-axis. Also, for a vector line with starting point (x1, y1) and final point (x2, y2), the direction is given by,

θ = tan-1 ((y2 – y1) / (x2 – x1))

References:
[1]https://www.cuemath.com/direction-of-a-vector-formula/#
[2]https://www.geeksforgeeks.org/direction-of-a-vector-formula/

7. What is the difference between a square matrix and a rectangular matrix?
A rectangular matrix is one in which the number of rows is not the same as the number of columns. Whereas, a square matrix has the same number of rows and columns.

Reference:
[1]https://www.vedantu.com/question-answer/a-matrix-that-is-not-a-square-matrix-is-called-a-class-12-maths-cbse-60cd78137dad4045be561679

8. What is a basis in linear algebra?
Basis denotes the set of vectors that can represent any vector in a given vector space through linear combinations. The condition is that the vectors have to be linearly independent and should lie within that span of vectors.
References:
[1]https://mikebeneschan.medium.com/how-to-understand-basis-linear-algebra-27a3bc759ae9
[2]https://people.math.carleton.ca/~kcheung/math/notes/MATH1107/wk09/09_basis_and_dimension.html

9. What is a linear transformation in linear algebra?
A linear transformation (also called a linear operator or map) is a function between two vector spaces that preserves vector addition and scalar multiplication, reflecting linearity properties. The field of the two vector spaces should be the same. The preservation of addition means that if two vectors are added and then their transformation is done then this result should be equal to the addition of individual transformations of the two vectors i.e., T(u+v)=T(u)+T(v). Here, T stands for transformation and u and v are the two vectors.
Also, preservation of scalar multiplication implies that when the transformation of the product of any vector v and scalar, c is taken then the result should be equal to the product of scalar, c, and transformation of vector, v.
References:
[1]https://brilliant.org/wiki/linear-transformations/#:~:text=A%20linear%20transformation%20is%20a,a%20linear%20operator%20or%20map.
[2]https://mathinsight.org/linear_transformation_definition_euclidean#:~:text=A%20linear%20transformation%20(or%20a,x)%2BT(y)

10. What is an eigenvector in linear algebra?

Eigen vector in linear algebra, is defined as that non-zero vector, v which when multiplied to a square matrix results in a scalar multiple of v which is called the eigenvalue. There are many applications of eigenvectors in machine learning and data science. These are often used to transform and reduce the dimensionality of data in various algorithms like PCA, SVD, and others.

References:

[1]https://colab.research.google.com/drive/1rp1uo_TFdb5n-ynTcwbWu05ttwTHIuiz?usp=sharing#scrollTo=73_lP8PjKz1F
[2]https://www.geeksforgeeks.org/eigen-values/

11. What is the gradient in machine learning?

Gradient measures how much change will be there in the output of a function if the inputs are tweaked a bit. In machine learning (ML), it refers to a vector containing the partial derivatives of a function with respect to each of its variables. It may also be referred to as the slope of a function and it points in the direction of the steepest increase of the function at a specific point. The more the value of the gradient, the steeper the slope will be and the model will learn faster. In ML, gradient is used to find the partial derivative of a function that generally denotes a loss or cost function which is used to find out the difference between the predicted values and the actual values. It refers to the measurement of change in all weights with regard to the change in error. Thus, in this way, the gradient, during the training of the ML model, guides the chosen optimization method to minimize the loss function by tweaking its weight and bias parameters in each iteration in order to enhance the performance. This is known as the gradient descent algorithm in which

Mathematically, if you have a function $J(\theta)$ representing the cost and θ is the vector of parameters, the gradient $\nabla J(\theta)$ is computed as:

$$\nabla J(\theta) = \left[\frac{\partial J}{\partial \theta_1}, \frac{\partial J}{\partial \theta_2}, \ldots, \frac{\partial J}{\partial \theta_n}\right]$$

where $\frac{\partial J}{\partial \theta_i}$ is the partial derivative of the cost with respect to the i-th parameter.

Reference:

[1] https://builtin.com/data-science/gradient-descent#

12. What is backpropagation in machine learning?

In machine learning (ML), the term backpropagation refers to the backward propagation or flow of errors." It is a supervised learning algorithm utilized in deep learning for training artificial neural networks to learn complicated patterns and relationships from data. Apart from that, it can also be applied to regression and classification problems.

It works by first passing the input data through the network to obtain the predicted output in the forward pass. Then, the error is calculated by comparison of the predicted and actual values. This calculated error is used in the backward pass to find out the weights for minimizing it in a gradient descent like process. After performing multiple iterations, the performance is improved.

Reference:

[1]https://www.techtarget.com/searchenterpriseai/definition/backpropagation-algorithm#:~:text=What%20is%20a%20backpropagation%20algorithm%20in%20machine%20learning%3F,values%20differ%20from%20actual%20output.

13. What is the concept of a derivative in calculus?

In calculus, a derivative denotes the instantaneous rate of change of a function with respect to another variable. Geometrically, it means the slope of the tangent to the graph of a function at a particular point. It provides information about the slope or steepness of a curve at a given point. The derivative is denoted by the symbol "dy/dx" or "f'(x)" and is used to solve optimization problems and to analyze the behavior of a function. Essentially, the derivative gives insights into how a function behaves locally, indicating how a small change in the input corresponds to a change in the output. In mathematical terms, it is expressed as the limit of the ratio of the change in the function's output to the change in its input, as the latter approaches zero. The derivative is fundamental for understanding motion, rates of change, and optimization in various mathematical and scientific applications.

Reference:

[1]https://www.cuemath.com/calculus/derivatives/#

14. How are partial derivatives used in machine learning?

The partial derivative is used to find the derivative of a function of different variables w.r.t. change in one of the variables. In ML, partial derivatives may be utilized for finding the minimum or maximum of a function, and in optimization algorithms used for training models. They enable the iterative refinement of model parameters to achieve better performance on specific tasks. For example, in the gradient descent algorithm, partial derivatives are employed to find out the gradient of the cost function w.r.t. the weight and basis parameters to minimize it. The partial derivatives of the cost function w.r.t. each parameter indicates the direction and magnitude of the steepest increase in the cost. By moving in the opposite direction of the gradient, the algorithm can iteratively update the parameters to minimize the cost.

Similarly, the backpropagation algorithm in neural networks that trains the network by minimizing the error between predicted and actual outputs, involves the calculation of partial derivatives of the loss w.r.t. the weights and biases throughout the network. These partial derivatives are used to update the weights and biases in the direction that minimizes the error, facilitating the learning process. Another example is feature engineering, which involves selecting and transforming input features. Partial derivatives can be used to assess the impact of each feature on the model's performance. This helps in understanding which features contribute most to the model's predictions.

Partial derivatives can also be employed in sensitivity analysis to understand how changes in input features affect the model's output. This is valuable for identifying influential features and understanding the robustness of the model. Additionally, many machine learning tasks involve optimizing a certain objective function, and partial derivatives provide information about the rate of change of the objective w.r.t. each variable. This information is crucial for making decisions and updating parameters to improve the model.

References:

[1]https://towardsdatascience.com/a-quick-introduction-to-derivatives-for-machine-learning-people-3cd913c5cf33
[2]https://machinelearningmastery.com/a-gentle-introduction-to-partial-derivatives-and-gradient-vectors/#:~:text=Partial%20derivatives%20and%20gradient%20vectors%20are%20used%20very%20often%20in,other%20classification%20and%20regression%20problems.
[3]https://machinelearningmastery.com/application-of-differentiations-in-neural-networks/

15. What is probability theory?

Probability theory is a branch of mathematics that deals with the study of uncertainty and randomness. It helps in understanding and quantifying the likelihood of events. It is defined as the number of favorable outcomes divided by the total number of possible outcomes of an event. For example, if a fair coin is tossed, what is the chance that it lands on the head?

In probability theory, events are defined as subsets of a sample space, which is the set of all possible outcomes of a random experiment. Probabilities are utilized to assign a numerical measure to the likelihood of each event occurring. Probability is always a number between 0 and 1, with 0 indicating that an event is impossible and 1 indicating that an event is certain.

Probability theory provides many useful tools for investigating randomness, such as probability distributions, random variables, and expected values.

References:
[1]https://www.cuemath.com/data/probability-theory/
[2]https://www.geeksforgeeks.org/probability-theory/

16. What are the primary components of probability theory?

Probability theory consists of several primary components that are used for handling uncertainty and randomness. There are four main components. Firstly, the **sample space** denotes the set of all possible outcomes in a given scenario. Secondly, **events** which can be defined as subsets of the sample space, represent the specific outcomes or combinations of outcomes. The core idea is to assign probabilities to these events, expressing how likely they are to occur.

Thirdly, a **probability measure** is a function that assigns a numerical value (i.e., probability) between 0 and 1 to each event, with 0 indicating impossibility and 1 indicating certainty.

Fourthly, **probability rules** that define how the calculation of probability occurs.For instance, the sum of all probabilities in the sample space is 1, and there are rules like the complement rule that help us calculate probabilities more efficiently.

Other components are as follows:

- A **random variable** is a variable that takes on values based on the outcomes of a random experiment. It maps the sample space to real numbers.
- The **probability distribution** of a random variable describes the probabilities associated with each possible value that the random variable can take. It can be represented by a probability mass function (PMF) for discrete random variables or a probability density function (PDF) for continuous random variables.
- For multiple random variables, the **joint probability distribution** describes the probability of simultaneous occurrences of specific values for each variable.
- **Conditional probability** measures the probability of an event A occurring given that another event B has occurred. It is a fundamental concept in probability theory.
  Independence:
- Events are considered **independent** if the occurrence of one event does not affect the probability of the other. Independence is a crucial concept in probability theory.

- The expected value, or **mean**, of a random variable is the average of all possible values, each weighted by its probability.
- **Variance** is a measure of the spread or dispersion of a random variable. It quantifies how much the values of a random variable deviate from its expected value.
- The **Law of Large Numbers** states that as the sample size increases, the sample mean converges to the expected value.
- The **Central Limit Theorem** states that the distribution of the sum (or average) of a large number of independent, identically distributed random variables approaches a normal distribution.

References:
[1]https://homepage.ntu.edu.tw/~ckuan/pdf/et01/et_Ch5.pdf
[2]https://www.slideshare.net/mairabc/basic-elements-of-probability-theory

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

Conditional probability is a type of probability that is defined as the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome. Conditional probability is calculated by multiplying the probability of the preceding event by the updated probability of the succeeding, or conditional, event. Conditional probability is often portrayed as the "probability of A given B," notated as P(A|B) and it can be calculated as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Here:

- $P(A|B)$ is the conditional probability of event A given that event B has occurred.
- $P(A \cap B)$ is the probability of both events A and B occurring simultaneously.
- $P(B)$ is the probability of event B occurring.

References:
[1]https://www.investopedia.com/terms/c/conditional_probability.asp#:~:text=Conditional%20probability%20is%20defined%20as,succeeding%2C%20or%20conditional%2C%20event.
[2]https://www.geeksforgeeks.org/conditional-probability/

18. What is Bayes theorem, and how is it used?

Bayes' Theorem, named after 18th-century British mathematician Thomas Bayes, is a mathematical formula for determining conditional probability. Conditional probability is the likelihood of an outcome occurring, based on a previous outcome having occurred in similar circumstances.

Bayes' theorem relies on incorporating prior probability distributions in order to generate posterior probabilities. Prior probability, in Bayesian statistical inference, is the probability of an event occurring before new data is collected. In other words, it represents the best rational assessment of the probability of a particular outcome based on current knowledge before an experiment is performed. Posterior probability is the revised probability of an event occurring after taking into consideration the new information. Posterior probability is calculated by updating the prior probability using Bayes' theorem. In statistical terms, the posterior probability is the probability of event A occurring given that event B has occurred.

$$\overset{\text{Posterior}}{\Downarrow \atop P(A|B)} = \frac{\overset{\text{Likelihood}}{\Downarrow \atop P(B|A)} * \overset{\text{Prior}}{\Downarrow \atop P(A)}}{\underset{\text{Evidence}}{P(B) \atop \Uparrow}}$$

Bayes' Theorem is particularly useful in Bayesian statistics, machine learning, and decision-making under uncertainty. Some example areas include medical diagnosis, spam filtering, fault diagnosis in engineering, document classification in Natural Language Processing, finance and investment, etc. In all these scenarios, Bayes' Theorem allows for the incorporation of new evidence or observations to refine or update prior probabilities, providing a more accurate and informed perspective in decision-making under uncertainty. It's a powerful tool for probabilistic reasoning in a wide range of applications.

References:

[1]https://www.investopedia.com/terms/b/bayes-theorem.asp#:~:text=Bayes%27%20Theorem%2C%20named%20after%2018th,having%20occurred%20in%20similar%20circumstances.
[2]https://www.freecodecamp.org/news/bayes-rule-explained/

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

A random variable is defined as the variable whose possible values are the outcomes of a random experiment. It is different from a normal variable as its values are related to probabliy and it performs the mapping of the outcomes of a random process to a numeric value. It can take any value. For example, when a die is rolled, random variable will be the outcome of the experiment i.e., the number that is obtained on the die. On the other hand, a regular variable is a deterministic variable denotes a fixed value which is not specified by uncertainty. Foe example, the number of students in a classroom, temperature of a substance, height of a specific individual, etc.
In summary, while both random variables and regular variables are used to represent quantities, a random variable introduces the element of uncertainty and probability, whereas a regular variable represents deterministic, specified values without randomness.

References:

[1]https://byjus.com/maths/random-variable/
[2]https://towardsdatascience.com/understanding-random-variable-a618a2e99b93#:~:text=After%20evaluating%2C%20x%3D5.,single%20value%20at%20a%20time.&text=Then%20X%20could%20be%200,might%20have%20a%20different%20probability.

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

The law of large numbers states that as the number of trials or experiments increases, the average of the results of those experiments will converge to the expected value. In other words, as the sample size increases, the average of the observed results will become more and more representative of the true value. In other words, if one performs a random experiment many times and calculates the average of the outcomes, the average will tend to get closer to the expected value as you conduct more trials.

The Law of Large Numbers has significant implications for probability theory because it provides a connection between theoretical probabilities and observed frequencies. It assures that, with a sufficiently large sample size, the empirical probabilities (based on observed outcomes) will closely match the true probabilities predicted by theoretical probability. This law is crucial for interpreting and applying probability in real-world situations, allowing us to make reliable predictions based on the long-term behavior of random phenomena.

For example, we want to estimate the average height of all the students in a school. If we measure the heights of just a few students (e.g. two percent), it is possible that the average height you calculate may not be very accurate. However, if we measure the heights of a larger group of students (let's say 35 percent), the average height you calculate is likely to be more representative of the true average height of all students in the school.

Another example of the law of large numbers is in gambling. If we flip a coin a few times, it is possible that you may get a string of heads or tails. However, if we flip the coin a large number of times, the proportion of heads and tails will tend to approach a ratio of 1:1, as the law of large numbers predicts.

Reference:
[1]https://builtin.com/data-science/law-of-large-numbers

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

The Central Limit Theorem (CLT) is a fundamental concept in probability theory and statistics that states that, under certain conditions, the distribution of the sum (or average) of a large number of independent, identically distributed random variables approaches a normal (Gaussian) distribution, regardless of the original distribution of the variables. Here, the random variables involved in the sum or average must be independent. This means that the occurrence of one event does not affect the occurrence of another. Also, the random variables should be drawn from the same probability distribution. They should have the same mean and the same standard deviation.

The CLT becomes more applicable as the sample size increases. Generally, a sample size of 30 or more is often considered sufficiently large for the CLT to be effective. It finds its applications in probability and statistics and some of the examples are given ahead. Firstly, it may be utilized to approximate the distribution of the sample mean or sum with a normal distribution, even if the original distribution of the individual random variables is not normal. Secondly, the CLT is also fundamental to many statistical inference techniques. For example, it justifies the usage of the normal distribution in constructing confidence intervals and performing hypothesis tests for population parameters. Thirdly, it is crucial in understanding the properties of sampling distributions. It allows statisticians to make probabilistic statements about the behavior of sample statistics. Fourthly, when dealing with large samples, the CLT is often invoked to make approximations and simplifications in statistical analyses. Fifthly, CLT is applied in quality control and process monitoring, where the distribution of sample means is used to assess the stability and performance of manufacturing processes. Sixthly, CLT may be used in risk management to model the distribution of portfolio returns, assuming that the returns are the sum of many independent and identically distributed random variables.Finally, CLT is used in simulation studies to generate random samples from a known distribution, providing a basis for making probabilistic statements about the behavior of certain statistics.

References:

[1]https://www.scribbr.com/statistics/central-limit-theorem/#:~:text=The%20central%20limit%20theorem%20says,the%20mean%20will%20be%20normal.

[2]https://www.investopedia.com/terms/c/central_limit_theorem.asp#:~:text=Why%20Is%20the%20Central%20Limit,easier%20statistical%20analysis%20and%20inference.

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

Discrete and continuous probability distributions are two fundamental types of probability distributions that describe the likelihood of different outcomes in a random experiment or process.

A discrete distribution is one in which the data can only take on certain values, for example, integers. A continuous distribution is one in which data can take on any value within a specified range (which may be infinite). For a discrete distribution, probabilities can be assigned to the values in the distribution – for example, "the probability that the web page will have 12 clicks in an hour is 0.15." In contrast, a continuous distribution has an infinite number of possible values, and the probability associated with any particular value of a continuous distribution is null. Therefore, continuous distributions are normally described in terms of probability density, which can be converted into the probability that a value will fall within a certain range.

In discrete distribution, the set of possible values is countable as compared to continuous distribution in which it is uncountable. For the representation of discrete distributions, the Probability Mass Function (PMF) is used while for continuous, the Probability Density Function (PDF) is used. In terms of graphical representation, discrete is often represented using histograms or bar graphs whereas continuous is represented by using smooth curves. An example of the former may be several students in a class while the latter may be the height of students in a class.

References:

[1]https://www.statistics.com/glossary/continuous-vs-discrete-distributions/#:~:text=A%20discrete%20distribution%20is%20one,(which%20may%20be%20infinite).

[2]https://support.minitab.com/en-us/minitab/21/help-and-how-to/probability-distributions-random-data-and-resampling-analyses/supporting-topics/basics/continuous-and-discrete-probability-distributions/

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

A measure of central tendency, also known as a measure of central location is a single value that attempts to describe or summarize a set of data by identifying the central position within that set of data. These are also classed as summary statistics. Firstly, mean (or average) is the most likely used measure of central tendency for symmetric distributions. It is calculated by summing up all the values in a dataset and dividing by the number of observations. Its disadvantage is that it is sensitive to extreme values (outliers). Also, the sum of deviations from the mean is always zero. Secondly, the median is the middle value in a dataset when the values are sorted in ascending or descending order. For an odd number of observations, the median is the middle value and for an even number of observations, the median is the average of the two middle values. It is less affected by extreme values compared to the mean and is appropriate for skewed distributions.

Thirdly, mode is the value that appears most frequently in a dataset. A dataset may have one mode (unimodal), more than one mode (multimodal), or no mode at all. It is appropriate for categorical data and discrete distributions.

References:

[1]https://byjus.com/maths/central-tendency/#:~:text=The%20three%20commonly%20used%20measures,of%20dispersion%20is%20the%20range.
[2]https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php

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

Percentiles and quartiles are quantiles. Quantiles are values that split sorted data or a probability distribution into equal parts. In general terms, a q-quantile divides sorted data into q parts. Percentiles and quartiles are statistical measures used to summarize and analyze the distribution of a dataset. They provide insights into the spread and central tendency of the data, allowing for a more detailed understanding of its characteristics. Percentiles and quartiles offer a standardized way to compare different datasets, especially in terms of central tendency and dispersion, and are widely used in decision-making.

A percentile is a measure indicating the relative standing of a particular value within a dataset. It represents the percentage of data points that fall below or equal to a given value. Percentiles help identify the spread of values in a dataset and locate specific data points in relation to the entire distribution. Common percentiles include the 25th (first quartile), 50th (median or second quartile), and 75th (third quartile) percentiles. Percentiles provide a detailed view of how individual values are distributed across the entire range. Percentiles allow for analysis of extreme values and their position in the dataset. In fields like healthcare and education, percentiles are used to assess performance or health indicators about a larger population.

Quartiles divide a dataset into four equal parts, each representing 25% of the data. There are three quartiles: Q1 (first quartile), Q2 (second quartile, or the median), and Q3 (third quartile). Quartiles provide a way to divide the data into segments, allowing for a quick assessment of the distribution's central tendency and spread. They are particularly useful for identifying potential outliers and understanding the distribution's skewness. Quartiles, especially the median, provide information about the central location of the data.

References:

[1] https://www.scribbr.com/statistics/quartiles-quantiles/#quantiles
[2] https://www.scribbr.com/statistics/quartiles-quantiles/

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

Detecting and treating outliers in a dataset is an important step in data analysis to ensure that extreme values do not unduly influence the results.
There are different ways in which outliers can be detected. Firstly, visual inspection may be done by the usage of box plots, histograms or scatter plots to find out any outlier that is deviating from the main distribution of the data points. Secondly, Z-score which measures how many standard deviations a data point is from the mean, can also be calculated for each data point to rule out outliers with high absolute Z-score values. Thirdly, by using Interquartile (IQR) method for skewed distributions, which involves finding the difference between the first (Q1) and the third quartile (Q3) and discarding any value that lies outside the range calculated as [Q1-1.5xIQR, Q3+1.5xIQR]. Various other methods such as percentiles, Mahalanobis Distance, DBSCAN (Density-Based Spatial Clustering of Applications with Noise, Local Outlier Factor (LOF), and One-Class SVM (Support Vector Machine), etc. may be used.

After detection of outliers, their handling is also important. Various methods may be used. Firstly, mathematical transformations may be appliedto the data, such as log transformations, to reduce the impact of extreme values. This can make the distribution more symmetric. Secondly, the extreme values may be set to a specified percentile, often the minimum or maximum values within a certain range. This helps limit the influence of outliers without entirely removing them. Thirdly, outliers may be replaced with imputed values, which can be derived from the mean, median, or a more sophisticated imputation method. Fourthly, statistical models that are robust to outliers may be used. Robust regression methods, like Huber regression, are less sensitive to extreme values. Fifthly, extreme values may be truncated by setting a threshold beyond which data points are considered outliers. This avoids completely removing them but limits their impact. Various other methods may also be used.

Reference:
[1]https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/
[2]https://towardsdatascience.com/outlier-detection-methods-in-machine-learning-1c8b7cca6cb8
[3]https://www.baeldung.com/cs/ml-outlier-detection-handling

26. How do you use the central limit theorem to approximate a discrete probability distribution?

The Central Limit Theorem allows us to approximate the distribution of the sample mean of a sufficiently large sample of independent, identically distributed discrete random variables as a normal distribution. This is achieved by calculating the mean and standard deviation of the original distribution and using them to define parameters for the approximate normal distribution. The larger the sample size, the better the approximation.

References:
[1] https://bookdown.org/peter_neal/math4081_notes/Sec_CLT.html#
[2]https://www.quora.com/What-is-the-central-limit-theorem-Is-it-valid-for-discrete-distributions#:~:text=The%20theorem%20is%20valid%20for,of%20successes%20in%20n%20trials.

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

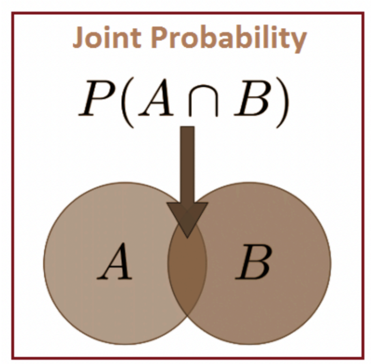
Goodness of fit measures how well observed data aligns with the expected distribution from a theoretical model. To test the goodness of fit for a discrete probability distribution, we can use methods like the Chi-Square Test, Kolmogorov-Smirnov Test, Cramér-von Mises Test, or Anderson-Darling Test. These tests compare observed and expected frequencies or cumulative distribution functions, providing statistical measures or p-values to assess how well the data fits the theoretical distribution. Visual methods like Q-Q plots or P-P plots can also aid in evaluating goodness of fit.

References:

[1]https://statisticsbyjim.com/hypothesis-testing/goodness-fit-tests-discrete-distributions/

28. What is a joint probability distribution?

The joint probability distribution describes the likelihood of multiple events occurring simultaneously. It is the probability of event Y occurring at the same time that event X occurs. Here, both events must be independent of one another, which means they aren't conditional or don't rely on each other. Joint probabilities can be visualized using Venn diagrams. It provides a way to model and analyze the probabilities associated with the joint occurrence of two or more random variables in a statistical or probabilistic system.



Formula and Calculation of Joint Probability

Notation for joint probability can take a few different forms. The following formula represents the probability of events intersection:

$$P\left(X \bigcap Y\right)$$

where:

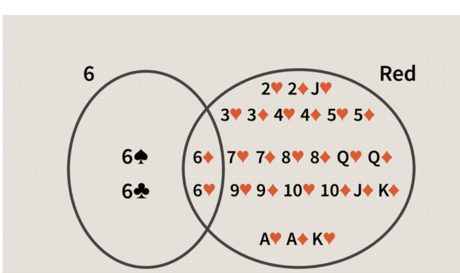
$X, Y =$ Two different events that intersect

$P(X \text{ and } Y), P(XY) =$ The joint probability of X and Y

For example, the joint probability of picking a card that is both red and 6 from a deck is P(6 ∩ red) = 2/52 = 1/26 since a deck of cards has two red sixes—the six of hearts and the six of diamonds. Because the events red and 6 are independent, we can also use the following formula to calculate the joint probability:

$$P(6 \cap red) = P(6) \times P(red) = 4/52 \times 26/52 = 1/26$$

The symbol "∩" in a joint probability is referred to as an intersection. The probability of event X and event Y happening is the same thing as the point where X and Y intersect. Therefore, the joint probability is also called the intersection of two or more events. A Venn diagram is perhaps the best visual tool to explain an intersection:



From the Venn above, the point where both circles overlap is the intersection, which has two observations: the six of hearts and the six of diamonds.

The joint probability distribution can be expressed in different ways based on the nature of the variable. In case of discrete variables, we can represent a joint probability mass function. For continuous variables, it can be represented as a joint cumulative distribution function or in terms of a joint probability density function.

Reference:

[1]https://byjus.com/maths/joint-probability/#:~:text=A%20joint%20probability%20distribution%20represents,a%20relationship%20between%20two%20variables.

[2]https://www.investopedia.com/terms/j/jointprobability.asp

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

While joint probability distribution describes the probabilities of different outcomes occurring simultaneously for two or more random variables, marginal probability distribution focuses on the probabilities of individual random variables, ignoring the values of other variables. Marginal probability, also referred to as unconditional probability is the probability of an event irrespective of the outcome of another variable. For example, the probability of drawing a black card from a deck. On the other hand, joint probability i.e., p(A and B) means the probability of event A and event B occurring together. For example, the probability that a card is a four and red =p(four and red) = p(A ∩ B)= 2/52=1/26. (There are two red fours in a deck of 52, the 4 of hearts and the 4 of diamonds).

While joint probability distribution involves calculating probabilities associated with specific combinations of values for all involved random variables, marginal probability distribution involves summing (or integrating) the joint probabilities over all possible values of the other variables to obtain the probability distribution for a single variable.

The joint distribution is inherently multidimensional, reflecting the joint behavior of the variables whereas the marginal distribution is unidimensional, focusing on the probabilities of individual variables without considering their joint behavior. Marginal probability distributions can be derived from the joint probability distribution by summing (or integrating) over the values of other variables. This process is known as marginalization. If you have a joint distribution for two variables X and Y, the marginal distribution for X is obtained by summing over all possible values of Y, and vice versa.

References:

[1]https://machinelearningmastery.com/joint-marginal-and-conditional-probability-for-machine-learning/#:~:text=Specifically%2C%20you%20learned%3A,presence%20of%20a%20second%20event.

[2]https://sites.nicholas.duke.edu/statsreview/jmc/

31. What is the covariance of a joint probability distribution?

Covariance is a measure of the degree to which two random variables X and Y change together. In joint probability distribution, covariance measures the change of the random variable X w.r.t. The random variable Y and it is defined as the expected value of the product of the deviations of X and Y from their respective means.
Mathematically, Cov(X,Y) = E[(X - E[X])(Y - E[Y])] where, Cov(X,Y)= covaraiance, E[.] refers to the expected value and X and Y are the random variables. If covariance is positive, high value of X would imply high value of Y and vice-versa. Similarly, if covariance is negative then high value of X would imply low value of Y and vice-versa. Also, if it is zero, then X and Y variables are independent of each other.

References:
[1]https://ocw.mit.edu/courses/18-05-introduction-to-probability-and-statistics-spring-2014/976210ba80799a72c8d0623d604aac06_MIT18_05S14_class7slides.pdf
[2]https://www.almabetter.com/bytes/tutorials/applied-statistics/joint-distribution-covariance-correlation

32. How do you determine if two random variables are independent based on their joint probability distribution?
Two events A and B are independent if their joint probability distribution is equal to the product of the marginal probability of A and the marginal probability of B.
P(A,B)=P(A and B)=P(A∩B)=P(A).P(B)
Intuitively, two random variables are independent if knowing the value of one of them does not change the probabilities for the other one. For example, when 2 dice are thrown, X is the result of the first coin toss (1 for heads, 0 for tails) and Y is the result of the second coin toss. Here, X and Y variables are independent because the probability of getting heads on the second toss is the same whether or not heads occurred on the first toss.

References:
[1]https://ocw.mit.edu/courses/18-05-introduction-to-probability-and-statistics-spring-2014/976210ba80799a72c8d0623d604aac06_MIT18_05S14_class7slides.pdf
[2]https://www.probabilitycourse.com/chapter3/3_1_4_independent_random_var.php#:~:text=Intuitively%2C%20two%20random%20variables%20X,%2C%20for%20all%20x%2Cy.

34. What is sampling in statistics, and why is it important?
In statistics, sampling refers to the process of selecting a subset of elements from a larger population in order to make inferences or draw conclusions about the entire population. The subset selected is called a sample, and the process of selecting the sample is known as the sampling process. For example, for researching the opinions of students in a university, a sample of 100 students can be taken and results can be generalized for all the students of that university. The use of samples allows researchers to conduct their studies with more manageable data and in a timely manner. There are various methods of sampling, including simple random sampling, stratified sampling, systematic sampling, and cluster sampling. Each method has its own advantages and is suitable for different types of studies and populations.

Sampling is important for several reasons:

1. Cost-Effective: It is generally impractical or way too expensive to collect data from an entire population. Sampling allows researchers to study a smaller subset of the population, saving time and resources.
2. Time Efficienct: Collecting data from an entire population can be time-consuming. Sampling enables researchers to gather information from a subset of the population, leading to faster data collection and analysis.
3. Practical: In some cases, it may be impossible to study an entire population due to logistical constraints. Sampling provides a practical way to study a representative subset of the population.
4. Inference: By studying a sample and making statistical inferences, researchers can draw conclusions about the population as a whole. If the sample is chosen correctly and is representative of the population, the findings can be generalized.
5. Accuracy: Well-designed sampling methods help ensure the accuracy and reliability of the data collected. Random sampling, for example, helps reduce bias and ensures that each member of the population has an equal chance of being included in the sample.

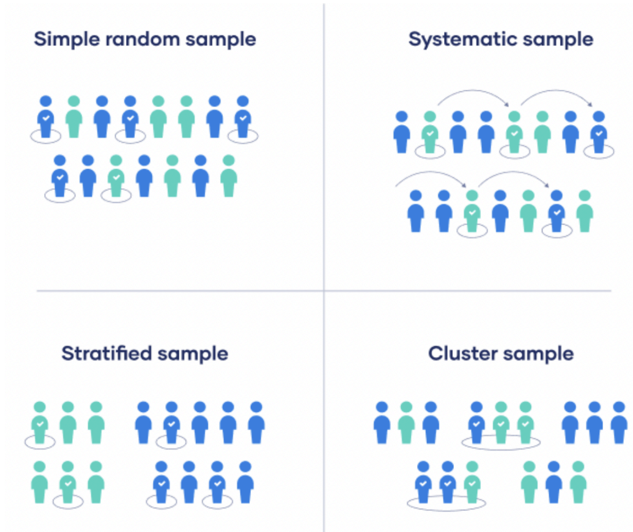
References:
[1]https://www.scribbr.com/frequently-asked-questions/what-is-sampling/
[2]https://www.cloudresearch.com/resources/guides/sampling/what-is-the-purpose-of-sampling-in-research/

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

There are two primary types of sampling methods widely used in research. First one is probability sampling which involves random selection, which allows making strong statistical inferences about the whole group. The second one is non-probability sampling which involves non-random selection based on convenience or other criteria, allowing easy collection of data.
Further, there are four main types of probability sampling techniques.

1. Simple Random Sampling: Every individual in the population has an equal chance of being included in the sample. In this, individuals are randomly selected without any specific pattern or order. Various tools like random number generators or other techniques that are based entirely on chance can be used. For example, to select a simple random sample of 1000 employees of a social media marketing company, we can assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

2. Stratified Sampling: The population is divided into subgroups or strata based on certain characteristics (e.g., age, gender), and then samples are randomly selected from each stratum. For example, For a company having 800 female and 200 male employees, in order to ensure that the sample reflects the gender balance of the company, sorting of population into two strata based on gender can be done. Then random sampling cane be performed on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

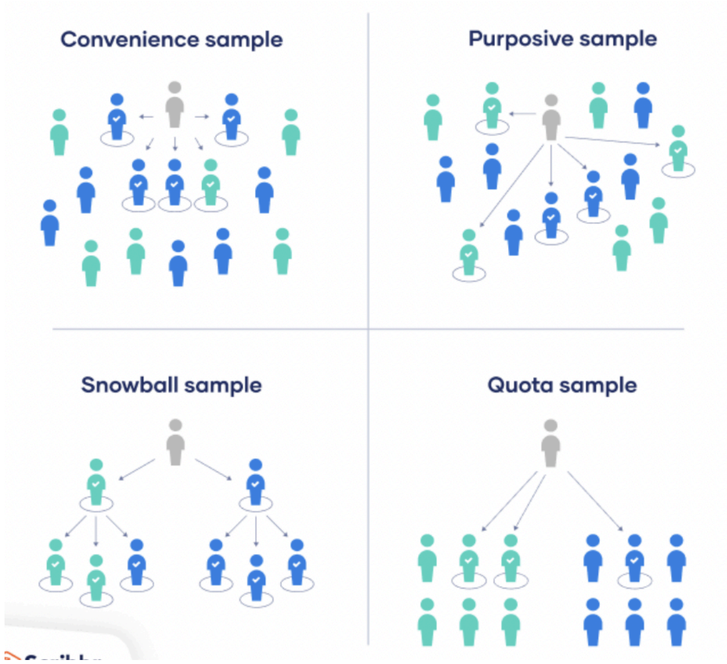


3. Systematic Sampling: It is similar to simple random sampling, but it is usually slightly easier to conduct. It involves Selecting every kth individual from the population after randomly selecting a starting point. For example, all employees of the company are listed in alphabetical order. From the first 10 numbers, random selection a starting point is done: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and one ends up with a sample of 100 people. While using this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

4. Cluster Sampling: The population is divided into clusters and then entire clusters are randomly selected to include in the sample. This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It's difficult to guarantee that the sampled clusters are really representative of the whole population. For example, there is a company that has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). It is not possible to travel to every office to collect the data, so random sampling is used to select 3 offices which are the clusters.

Similarly, there are four types of non-probability sampling methods:
This type of sample is easier and cheaper to access, but it has a higher risk of sampling bias. That means the inferences made about the population are weaker than with probability samples, and the conclusions may be more limited. If a non-probability sample is used, it should be made it as

representative of the population as possible. Non-probability sampling techniques are often used in exploratory and qualitative research. In these types of research, the aim is not to test a hypothesis about a broad population, but to develop an initial understanding of a small or under-researched population.



1.Convenience Sampling: It involves choosing individuals who are easiest to reach or include in the study rather than using a random or systematic method. This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can't produce generalizable results. Convenience samples are at risk for both sampling bias and selection bias. For example, while researching opinions about student support services in a university, the research student, after every class, a fellow student is asked to complete a survey on the topic. This is a convenient way to gather data, but as only students are surveyed taking the same classes as the research student at the same level, the sample is not representative of all the students at their university.

2.Snowball Sampling: It involves identifying initial participants and then asking them to refer others to be part of the sample. Starting with a small group, expansion of the sample is done through referrals. If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to "snowballs" as you get in contact with more people. The downside here is also representativeness, as you have no way of knowing how representative your sample is due to the reliance on participants recruiting others. This can lead to sampling bias. For example, you are researching experiences of homelessness in your city. Since there is no list of all homeless people in the city, probability sampling isn't possible. You meet one person who agrees to participate in the research, and she puts you in contact with other homeless people that she knows in the area.

3.Quota Sampling: It involves setting quotas for certain characteristics (e.g., age, gender) and then non-randomly selecting individuals to meet those quotas. Quota sampling relies on the non-random selection of a predetermined number or proportion of units. This is called a quota. You first divide the population into mutually exclusive subgroups (called strata) and then recruit sample units until you reach your quota. These units share specific characteristics, determined by you prior to forming your strata. The aim of quota sampling is to control what or who makes up your sample.For example, you want to gauge consumer interest in a new produce delivery service in Boston, focused on dietary preferences. You divide the population into meat eaters, vegetarians, and vegans, drawing a sample of 1000 people. Since the company wants to cater to all consumers, you set a quota of 200 people for each dietary group. In this way, all dietary preferences are equally represented in your research, and you can easily compare these groups.You continue recruiting until you reach the quota of 200 participants for each subgroup.

4.Purposive Sampling: It involves selecting individuals based on specific criteria or characteristics relevant to the research question. This type of sampling, also known as judgement sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research. It is often used in qualitative research, where the researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inferences, or where the population is very small and specific. An effective purposive sample must have clear criteria and rationale for inclusion. Always make sure to describe your inclusion and exclusion criteria and beware of observer bias affecting your arguments. Example: You want to know more about the opinions and experiences of disabled students at your university, so you purposefully select a number of students with different support needs in order to gather a varied range of data on their experiences with student services.

References:
[1]https://www.scribbr.com/methodology/sampling-methods/
[2]https://www.questionpro.com/blog/types-of-sampling-for-social-research/

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

The Central Limit Theorem (CLT) states that, under certain conditions, the distribution of the sum or average of a large number of independent, identically distributed random variables approaches a normal distribution, regardless of the original distribution. It is crucial in statistical inference because it allows us to make probabilistic statements about sample means, even when the underlying population distribution is unknown or not normal. The CLT is foundational for hypothesis testing, confidence intervals, and other statistical methods, enabling reliable inferences in various fields.

References:
[1]https://www.investopedia.com/terms/c/central_limit_theorem.asp
[2]https://byjus.com/jee/central-limit-theorem/

41. What is the difference between correlation and causation?

Correlation and causation are two concepts in statistics and research that describe different types of relationships between variables. While correlation refers to a statistical measure that quantifies the degree to which two variables change together, It does not imply a cause-and-effect relationship but rather measures the strength and direction of the linear association between variables. On the other hand, causation refers to a cause-and-effect relationship between two variables. If changes in one variable lead to changes in another, there is a causal relationship. For example, if there is a high positive correlation between the number of ice cream cones sold and the number of drowning incidents at a beach, it does not imply that buying ice cream causes drowning. Instead, both variables may be influenced by a third variable, such as temperature. Also, as a second example, if we consider a study that finds a correlation between studying more hours and achieving higher grades, it does not necessarily mean that studying more causes higher grades. Other factors, such as motivation or prior knowledge, could influence both studying behavior and academic performance.

Reference:

[1]https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/correlation-and-causation#:~:text=A%20correlation%20between%20variables%2C%20however,relationship%20between%20the%20two%20events.

42. How is a confidence interval defined in statistics?

When an estimate is made in statistics, whether it is a summary statistic or a test statistic, there is always uncertainty around that estimate because the number is based on a sample of the population being studied. The confidence interval is the range of values that are expected by the estimate to fall between a certain percentage of the time if the experiment is run again or the population is re-sampled in the same way. Hence, a confidence interval, refers to the probability that a population parameter will fall between a set of values for a certain proportion of times. Analysts often use confidence intervals that contain either 95% or 99% of expected observations. Thus, if a point estimate is generated from a statistical model of 10.00 with a 95% confidence interval of 9.50 - 10.50, it means we are 95% confident that the true value falls within that range. Thus, the confidence level is the percentage of times you expect to reproduce an estimate between the upper and lower bounds of the confidence interval.

References:

[1]https://www.scribbr.com/statistics/confidence-interval/#:~:text=The%20confidence%20interval%20is%20the,population%20in%20the%20same%20way.
[2]https://www.investopedia.com/terms/c/confidenceinterval.a

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

The confidence level represents the percentage of times one expects to reproduce an estimate between the upper and lower bounds of the confidence interval, and is set by the alpha value.
In statistics, a confidence interval is a range of values that is used to estimate the true value of an unknown parameter with a certain level of confidence. The confidence level associated with a confidence interval represents the probability that the interval contains the true value of the parameter. The confidence level is a percentage or a probability that is associated with the confidence interval. Commonly used confidence levels are 90%, 95%, and 99%. For example, suppose we calculate a 95% confidence interval for the average height of a population to be between 65 inches and 75 inches. This does not mean that there is a 95% chance that the true average height falls within this range. Instead, it means that if we were to construct many confidence intervals from different samples, about 95% of them would contain the true average height. For example, what a 95 percent confidence level is implying is that if the poll or survey were repeated over and over again, the results would match the results from the actual population 95 percent of the time. Also, higher confidence levels (e.g., 99%) result in wider confidence

intervals. This is because a higher level of confidence implies a greater degree of certainty, requiring a larger interval to capture the true parameter with the desired probability.
Choosing a confidence level involves a trade-off between precision and certainty. A higher confidence level provides greater certainty but results in wider intervals, which may be less precise.

Reference:

[1]https://www.indeed.com/career-advice/career-development/confidence-level-versus-confidence-interval

[2]https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/confidence-level/

48. What are the key elements to consider when designing an experiment?

When designing an experiment, key elements to consider include the research question, variables, experimental units, randomization, control group, replication, blocking, blinding, statistical analysis, ethical considerations, unbiasedness of the experiment, validity, reliability, practical constraints, and measurement techniques. These elements collectively contribute to a well-designed experiment that produces meaningful and reliable results.

References:

[1]https://nc3rs.org.uk/3rs-resources/key-elements-well-designed-experiment

[2]https://www.voxco.com/blog/what-are-the-5-components-of-experimental-design/

49. How can sample size determination affect experiment design?

The choice of the sample size has a large effect on the validity, reliability and generalizability of the experiment design and is its determination is very important.

If the sample size is large, the results of the experiment will be more accurate, precise and there will be less variability in the sample estimates. A large sample size can reduce the risk of Type II errors, where the study fails to detect a true effect. A larger sample increases the likelihood of detecting effects that are present.

However, if a sample size is too large, beyond requirement, the increase in accuracy will not be that much and it would cause unnecessary wastage of resources such as cost, participants, time, etc. for collecting data and it might cause inconvenience to the population too from whom the sample is being taken. For example, if the sample is taken on patients in a hospital, taking an etremely large sample would be unethical. In this regard, inconvenience to patients refers to the time that they spend in clinical assessments and to the psychological and physical discomfort that they experience in assessments such as interviews, blood sampling, and other procedures.

In contrast, a sample that is smaller than necessary would have insufficient statistical power to answer the primary research question, and a statistically nonsignificant result could merely be because of inadequate sample size (Type 2 or false negative error). Thus, a small sample could result in the patients in the study being inconvenienced with no benefit to future patients or to science. This is also unethical. An extremely small sample size increases the risk of Type I errors, where the study incorrectly identifies an effect that does not exist. This is because the likelihood of obtaining a statistically significant result by chance is higher with a small sample.

However, there are certain advanatges to the experiment design if the sample is sufficiently small. An experiment with a small sample can provide more believable results than those on a large sample with uncontrolled confounders. Small samples have a tremendous advantage as highly sophisticated and accurate measurements can be made with all the precautions in place. The measurement errors and biases can be easily controlled and can be easily identified in a small sample. The aggregation errors that occur due to the combining of small and large values are less likely with small samples. Small samples give quick results, can be carried out in one center without the hassles of multicenter studies, and are easy to get the ethical committee approval. They may require exact methods of statistical analysis that can help in reaching more valid conclusions. The smaller the sample is, the less likely it is to be confident that the results reflect the true population parameter.

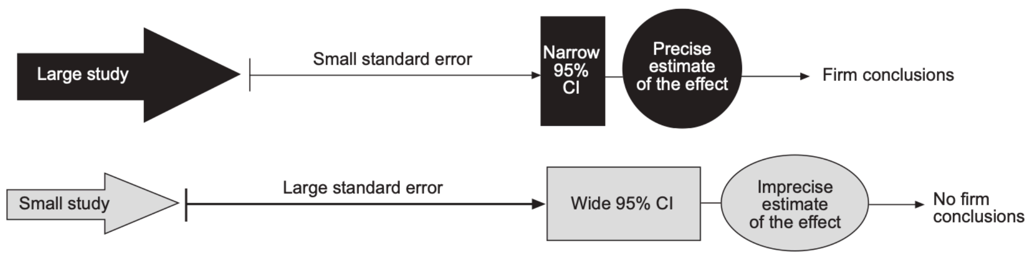


Figure: Schematic diagram showing how study size can influence conclusions (CI: confidence interval) [4]

Hence, researchers need to balance the desire for statistical precision with ethical considerations, particularly in fields where experimentation involves human subjects. Determining an appropriate sample size ensures that the study is ethically conducted.

References:

[1]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6970301/#:~:text=A%20sample%20that%20is%20larger,in%20recruiting%20the%20extra%20patients.
[2]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8706541/
[3]https://blog.statsols.com/why-is-sample-size-important/
[4]https://erj.ersjournals.com/content/erj/32/5/1141.full.pdf

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

Mitigating sources of bias is crucial in experimental design to ensure that the study accurately reflects the effects of the manipulated variables. Here are some strategies to mitigate potential sources of bias in experiment design:

1. Randomization: Randomly assign participants to different experimental conditions or treatments. This helps ensure that each group is, on average, similar in terms of both observed and unobserved characteristics, reducing selection bias. When possible, use random sampling to select participants from the target population. Random selection enhances the generalizability of the findings to the larger population.

2. Counterbalancing: If the order of presenting conditions might influence results, counterbalance the order of conditions across participants to control for order effects.

3. Control Groups: Include control groups to provide a baseline for comparison. The control group should be treated similarly to the experimental group but without the manipulation, allowing for a more accurate assessment of the treatment effect.

4. Blinding: Implement double-blind procedures where both participants and experimenters are unaware of the treatment conditions. This helps prevent bias resulting from participants' expectations or experimenters' influence. If double-blind procedures are not feasible, consider single-blind procedures where participants are unaware of their treatment condition.

5. Matching: Pair participants who are similar on relevant characteristics before random assignment to different conditions. Pair matching helps control for individual differences and reduces the risk of confounding variables.

6. Minimizing Demand Characteristics: If possible, use deceptive techniques to prevent participants from guessing the true purpose of the study, reducing demand characteristics that might bias their behavior.

7. Careful Measurement: Use reliable and valid measures to assess outcomes. Ensure that measurement tools accurately capture the variables of interest. If applicable, include pre-post measures to control for individual differences and changes over time.

8. Sampling Techniques: If certain subgroups are of particular interest, use stratified sampling to ensure representation from each subgroup in the sample.

9. Analytical Strategies: Use statistical control techniques, such as analysis of covariance (ANCOVA), to control for the influence of potential confounding variables. If matching is used during the design phase, consider matching in the analysis to account for potential biases.

10. Transparent Reporting: Clearly document the experimental design, procedures, and statistical analyses in a transparent manner. This allows for the replication of the study and enhances the trustworthiness of the findings.

11. Peer Review: Submit the research plan and findings for peer review to receive feedback from experts in the field. Peer review helps identify potential biases and enhances the rigor of the study.

12. Ethical Considerations: Ensure that the experiment is conducted ethically, adhering to ethical guidelines and standards. Ethical conduct contributes to the validity and reliability of the study.

By employing these strategies, researchers can minimize potential sources of bias in experimental design, leading to more robust and credible findings. It's important to carefully consider these strategies in the planning phase to enhance the internal and external validity of the experiment.

References:

[1]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2917255/
[2]https://www.indeed.com/career-advice/career-development/how-to-avoid-researcher-bias
[3]https://www.surveymonkey.com/mp/dont-let-opinions-sneak-survey-4-ways-avoid-researcher-bias/
[4]https://www.scribbr.com/category/research-bias/

51. What is the geometric interpretation of the dot product?

The dot product of two vectors gives the measure or amount of similarity in geometric space between them. This can be found out by using the concept of projections. The dot product of two vectors A and B can be found out by multiplying their magnitudes with the cosine of the angle between them. Also, the projection of vector B on vector a is given as the product of magnitude of vector B and the cosine of the angle between them. Thus, we can say that the dot product of two vectors is equal to the product of magnitude of first vector and the projection of the second vector on the first vector. This is the geometric interpretation of the dot product.

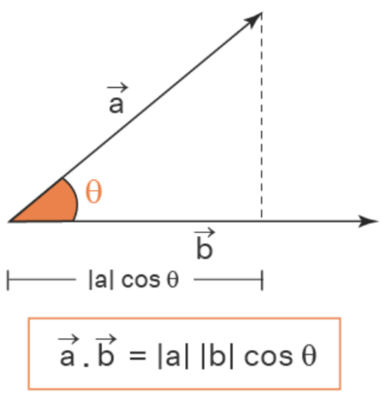


Figure: Dot product of vector a and vector b [2]

If the value of dot product is greater than zero (i.e., it is positive), it implies that the two vectors are closely aligned to each other. On the other hand, there is no alignment between them if the dot product is zero and there is 'opposite' alignment if the dot product is negative. This is the geometric interpretation of dot products of vectors which involves the concept of projection and angle between vectors.

Reference:

[1] https://www.youtube.com/watch?v=ZXChzVtIXz0
[2]https://www.cuemath.com/algebra/product-of-vectors/#

52. What is the geometric interpretation of the cross-product?

The geometrical interpretation of a function essentially means it's plot on graph. The cross product has a simple geometrical interpretation. The magnitude of the cross product of two vectors is the area of the parallelogram with the two vectors as adjacent sides, and the direction is that perpendicular to both the vectors (where the exact direction is decided by the right-hand rule).

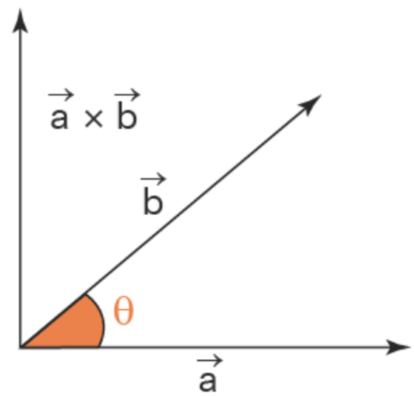


Figure: Cross product of vector a and vector b

$$\vec{a} \times \vec{b} = |a||b| \sin(\theta)\hat{n}$$

Here, n hat gives the direction vector. In physics, the cross product is particularly useful when there is some rotating object (a rigid body, a circulating fluid, a magnetic or electric field, etc.) with an axis of rotation - then there is usually an important associated quantity (torque, curl, etc.) defined using the cross product, in such a way that the direction is the same as that of the axis of rotation.The cross product is a versatile mathematical operation with physical significance. It allows us to understand and quantify rotational aspects of motion, making it a valuable tool in various fields, including physics and engineering. The "twist" or rotational influence mentioned refers to the way in which the cross product describes the rotational aspects of vectors in three-dimensional space.

References:
[1]https://byjus.com/maths/cross-product/
[2]https://www.khanacademy.org/math/multivariable-calculus/thinking-about-multivariable-function/x786f2022:vectors-and-matrices/a/cross-products-mvc

53. How are optimization algorithms with calculus used in training deep learning models?

Optimization algorithms with calculus play a crucial role in training deep learning models. Firstly, the training process involves adjusting the parameters (weights and biases) of a neural network to minimize a cost or loss function. This optimization task is achieved using iterative algorithms that leverage calculus concepts, such as gradients and derivatives. A cost function (also known as a loss function) is defined to measure the difference between the predicted output of the neural network and the actual target output. The goal is to minimize this cost by using the gradient descent algorithm.The gradient of the cost function with respect to the model parameters is computed. The gradient points in the direction of the steepest increase in the cost. The negative gradient points in the direction of the steepest decrease. Parameters are adjusted in the opposite direction to the gradient to reduce the cost. Mathematically, the parameters of the model i.e., weights (W) and bias (b) are re-calculated in every iteration by subtracting from their gradient multiplied by the learning rate (alpha).

$$
\begin{aligned}
&\text{Repeat } \{ \\
&W = W - \alpha \frac{\partial}{\partial W} J(W) \\
&b = b - \alpha \frac{\partial}{\partial b} J(b) \\
&\}
\end{aligned}
$$

Parameters are adjusted in the opposite direction of the gradient to reduce the cost. The update rule is defined as: $\theta = \theta - \alpha \cdot \nabla J(\theta)$

- θ represents the model parameters (weights and biases).
- α is the learning rate, a hyperparameter that determines the step size in the parameter space.
- $\nabla J(\theta)$ is the gradient of the cost function with respect to the parameters.

2. Stochastic Gradient Descent (SGD) and Mini-Batch Gradient Descent: In SGD, instead of using the entire training dataset to compute the gradient, a single randomly chosen data point (or a small batch) is used. This reduces computation time and helps avoid local minima. In Mini-Batch Gradient Descent, acompromise between SGD and batch gradient descent, where a small random subset (mini-batch) of the training data is used to compute the gradient.

3. Optimization Algorithms with Momentum and Adaptive Learning Rates: Momentum is introduced to accelerate convergence and overcome oscillations in the parameter space. It accumulates a moving average of past gradients to smooth out updates. Algorithms like AdaGrad, RMSprop, and Adam dynamically adjust the learning rates for each parameter based on the historical information of the gradients. This helps in dealing with varying magnitudes of gradients.

4. Backpropagation: The optimization algorithm, combined with the chain rule of calculus, is used in the backpropagation algorithm to efficiently compute gradients for each layer of the neural network. This enables the systematic update of weights during training.

5. Automatic Differentiation: Libraries like TensorFlow and PyTorch leverage automatic differentiation to automatically compute gradients. This simplifies the process of implementing complex neural network architectures and training procedures.

6. Regularization: Optimization algorithms are often extended with regularization techniques, such as L1 or L2 regularization, to prevent overfitting and improve the generalization of the model.

References:

[1]https://machinelearningmastery.com/calculus-in-machine-learning-why-it-works/
[2]https://medium.com/@chabavictor7/math-for-deep-learning-7c45c76642e8
[3]https://medium.com/analytics-vidhya/optimization-algorithms-for-deep-learning-1f1a2bd4c46b

54. What are observational and experimental data in statistics?

Depending on how data is collected, it can be classified as observational or experimental. Observational data (also called found data) is collected through observation i.e., it refers to anything that can be heard or seen. Observational data is the easiest to collect.
For example, observational data might be things like website data (visits, clicks, time spent on site, etc.), sales, emails, number of calls, etc. Observational data can also be called found data. It is a byproduct of other things. For example, a social media post is data by itself, but if we could also collect the relationships found in that post (with other users, pages, etc.), we would also have found data.

On the other hand, experimental data is collected by doing experiments through the scientific method with a prescribed methodology. This means that experimental data is not passively collected. Experimental data is collected for a specific purpose or question. It is also harder and more expensive to collect. Common mistakes with experimental data include not controlling for confounding variables, small sample sizes and using the wrong statistical tests.

While observational data has no control over variables i.e., observes existing conditions, the experimental data is obtained in a controlled environment with intentional manipulation of variables. The observational Data can suggest associations but does not establish causation. However, experimental data can establish causation due to controlled conditions and manipulation of variables.

In observational data, there is no random assignment; subjects are observed in their natural state but experimental data often involves random assignment of subjects to different groups. The observational data is often used when experimental manipulation is not feasible or ethical. On the other hand, experimental data is used to test causal relationships and establish the effects of interventions. Observational studies are valuable for exploring associations and generating hypotheses, while experimental studies are essential for testing causal relationships and making informed interventions. The choice between observational and experimental approaches depends on the research question, ethical considerations, and practical constraints.

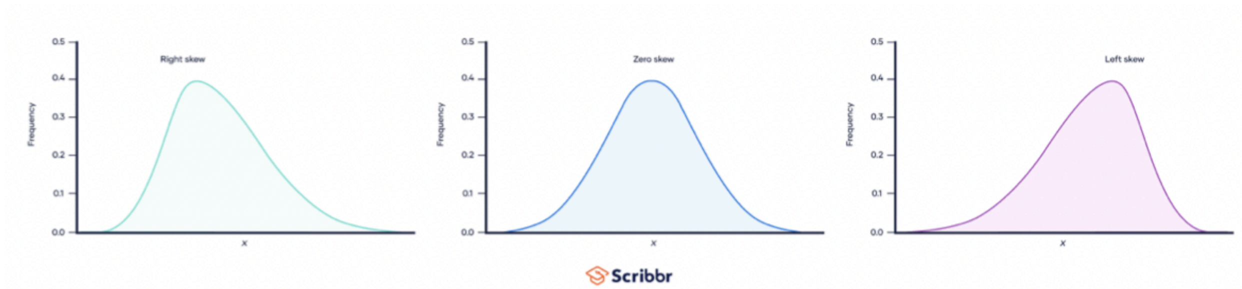
Reference:

[1]https://www.linkedin.com/pulse/observational-vs-experimental-data-whats-difference-rendón-ruiz/

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

Skewness is a measure of the asymmetry of a distribution. A distribution is asymmetrical when its left and right side are not mirror images.

A distribution can have right (or positive), left (or negative), or zero skewness. A right-skewed distribution is longer on the right side of its peak, and a left-skewed distribution is longer on the left side of its peak.



The mean of a right-skewed distribution is almost always greater than its median. That's because extreme values (the values in the tail) affect the mean more than the median. The mean of a left-skewed distribution is almost always less than its median.

Reference:

[1]https://www.scribbr.com/statistics/skewness/#:~:text=called%20positive%20skew).-,A%20right%2Dskewed%20distribution%20is%20longer%20on%20the%20right%20side,right%20sides%20are%20mirror%20images.

58. What is kurtosis?

Kurtosis is a measure of the tailedness of a distribution. There are three types of Kurtosis. First one is Normal Kurtosis which refers to distributions with medium kurtosis (medium tails) are mesokurtic. A mesokurtic distribution has a kurtosis value of 0. Second type is Excess Kurtosis (< 0) which refers to distributions with low kurtosis (thin tails) that are platykurtic. If the distribution has lighter tails and a flatter peak than the normal distribution, it is termed platykurtic. Platykurtic distributions have negative kurtosis. The third category is Excess Kurtosis(> 0) which comprises distributions with high kurtosis (fat tails) that are leptokurtic. If the distribution has heavier tails and a sharper peak than the normal distribution, it is termed leptokurtic. Leptokurtic distributions have positive kurtosis. Here, tails are the tapering ends on either side of a distribution. They represent the probability or frequency of values that are extremely high or low compared to the mean. In other words, tails represent how often outliers occur.

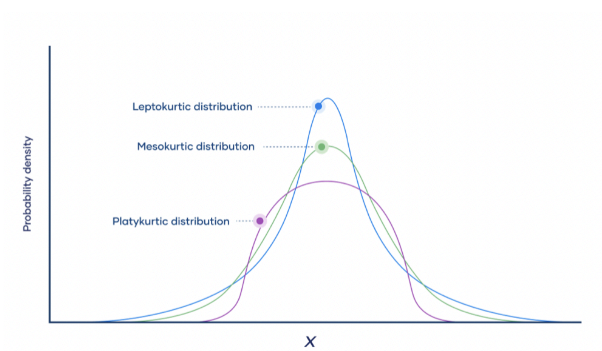


Figure: Types of kurtosis

References:

[1]https://www.scribbr.com/statistics/kurtosis/#mesokurtic
[2]https://www.scribbr.com/statistics/kurtosis/#platykurtic
[3]https://www.scribbr.com/statistics/kurtosis/#leptokurtic

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

Descriptive statistics summarize and describe the main features of a dataset, providing insights into its central tendency, dispersion, and shape. It focuses on the entire population and not on a sample of it.

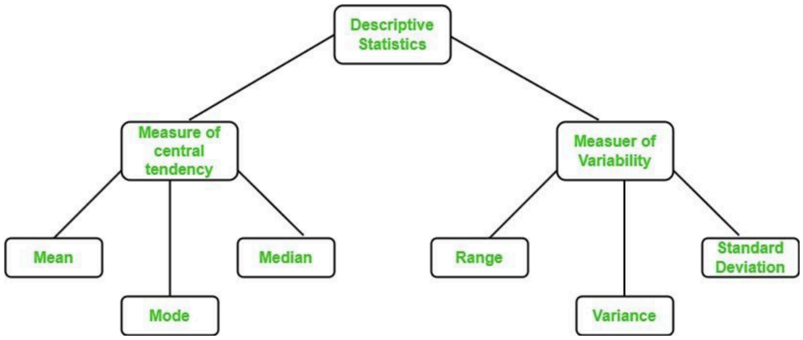


Figure: Descriptive statistics provides measures of central tendency and variability. (Eg.: Mean, median, mode, standard deviation, range, frequency tables) [1]

Inferential statistics, on the other hand, draw conclusions or make predictions about a population based on a sample of data, using probability theory.

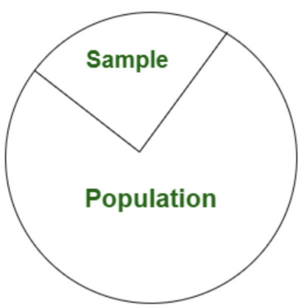


Figure: Inferential statistics estimates parameters, tests hypotheses, and determines the level of confidence or significance in the results on a sample of the population. (Eg.: Hypothesis testing, confidence intervals, regression analysis, ANOVA (analysis of variance), chi-square tests, t-tests, etc.) [1]

Descriptive statistics aim to present, analyze, interpret and organize data, while inferential statistics aims to generalize findings to a larger population, make predictions, test hypotheses, evaluate relationships, and support decision-making. While descriptive statistics is used for describing a situation, inferential statistics is used to explain the chance of occurrence of an event. Descriptive statistics explains already known data and is limited to a sample or population having a small size. Inferential statistics attempts to reach the conclusion about the population. While descriptive statistics can be achieved with the help of charts, graphs, tables, etc., inferential statistics can be achieved by probability.

References:

[1]https://www.geeksforgeeks.org/difference-between-descriptive-and-inferential-statistics/

[2]https://onlinedegrees.bradley.edu/blog/whats-the-difference-between-descriptive-and-inferential-statistics/#:~:text=Essentially%2C%20descriptive%20statistics%20state%20facts,make%20predictions%20about%20larger%20populations.
[3]https://www.simplilearn.com/difference-between-descriptive-inferential-statistics-article

62. 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?

A high correlation between sleep duration and productive work suggests a statistical association, but it doesn't imply causation. It could indicate a positive relationship, but factors like individual differences or third variables may contribute. Further research, such as controlled experiments, is needed to explore causation and underlying mechanisms.

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

A sampling error occurs when the sample used in the study does not represent the entire population. Although sampling errors occur frequently, researchers always include a margin of error in their conclusions as a matter of statistical practice. The margin of error is the amount allowed for a miscalculation to represent the difference between the sample and the actual population. Sampling is a type of analysis where a small sample of observations is chosen from a larger population. The selection bias process can produce both sampling errors and non-sampling errors. Sampling errors occur due to a disparity in the representativeness of the respondents. It majorly happens when the researcher does not plan his sample carefully. Sampling errors are caused because the sample size is small and is inadequate to capture the population behavior accurately. It can be reduced by increasing the sample size.

These sampling errors can be controlled and eliminated by creating a careful sample design, having a large enough sample to reflect the entire population, or using an online sample or survey audiences to collect responses.

References:

[1]https://byjus.com/question-answer/sampling-error-can-be-reduced-by-recording-data-more-carefullyreducing-the-sampling-biasincreasing-the-sample/#
[2]https://www.questionpro.com/blog/sampling-error/

80. What is an inlier?

An inlier is a data value that lies in the interior of a statistical distribution and is in error. An inlier is an observation lying within the general distribution of other observed values, generally does not perturb the results but is nevertheless non-conforming and unusual. Inliers are difficult to distinguish from good data values. For example, a simple example of an inlier might be a value in a record reported in the wrong units, say degrees Fahrenheit instead of degrees Celsius

Reference:

[1]https://community.datarobot.com/t5/product-support/what-is-an-inlier/td-p/11855