**End Course Summative Assignment**.

1. **What is a vector in mathematics?**

In mathematics, a vector is a quantity that has both magnitude and direction. Vectors are used to represent physical quantities such as displacement, velocity, and force. They can be visualized as arrows in a coordinate system, where the length of the arrow represents the magnitude, and the direction of the arrow indicates the direction of the quantity.

1. **How is a vector different from a scalar?**

A vector and a scalar are different types of quantities used in mathematics and physics, distinguished primarily by their attributes and how they are used to describe physical phenomena.

**Scalar:**

* **Definition**: A scalar is a quantity that has only magnitude (size or numerical value) and no direction.

**Vector:**

* **Definition**: A vector is a quantity that has both magnitude (size or numerical value) and direction.

1. **How can vectors be multiplied by a scalar?**

Multiplying a vector by a scalar involves multiplying each component of the vector by the scalar.

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

The direction of a vector can be determined using its components, often represented in terms of angles or as a unit vector.

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

The difference between a square matrix and a rectangular matrix lies in their dimensions and certain properties:

**Square Matrix:**

**Definition**: A square matrix is a matrix where the number of rows is equal to the number of columns.

**Dimensions**: If a square matrix has n rows, it also has n columns.

**Shape**: It has equal dimensions, such as n×n (e.g., 2x2, 3x3, etc.).

**Properties**:

* + Inverse: A square matrix can have an inverse if its determinant is non-zero. entries contains such.

1. **What is a basis in linear algebra?**

In linear algebra, a basis is a fundamental concept that describes the "building blocks" or foundation upon which vectors in a vector space can be uniquely represented. Here’s a concise explanation:

**Basis Definition:**

* **Definition**: A basis of a vector space VVV is a set of vectors that are linearly independent and span VVV.

1. **What is a linear transformation in linear algebra?**

In linear algebra, a linear transformation (or linear map) between two vector spaces VVV and WWW is a function T:V→W that preserves vector addition and scalar multiplication.

1. **What is an eigenvector in linear algebra?**

In linear algebra, an eigenvector is a non-zero vector that, when multiplied by a given square matrix, yields a new vector that is simply a scaled version of the original vector. This scaling factor is called the eigenvalue associated with that eigenvector.

1. **What is the gradient in machine learning?**

In machine learning, the gradient typically refers to the gradient of a loss function with respect to the model parameters.

1. **What is backpropagation in machine learning?**

Backpropagation is a fundamental algorithm used in the training of artificial neural networks, especially in the context of supervised learning tasks. It is essential for efficiently computing the gradients of a loss function with respect to the weights of the neural network, which allows for effective optimization using gradient-based methods like stochastic gradient descent (SGD). Here’s a concise explanation of backpropagation.

1. **What is the concept of a derivative in calculus?**

In calculus, the derivative is a fundamental concept that measures how a function changes as its input changes. It provides the rate of change of the function at a particular point and plays a crucial role in understanding the behaviour of functions, optimization, and many applications in science and engineering.

1. **How are partial derivatives used in machine learning?**

Partial derivatives play a crucial role in machine learning, especially in the context of optimizing complex models with multiple parameters.

1. **What is probability theory?**

Probability theory is a branch of mathematics that deals with quantifying uncertainty and analysing random phenomena. It provides a framework for understanding and predicting outcomes in situations where randomness, chance, or uncertainty is present.

1. **What is Bayes theorem, and how is it used?**
2. **Machine Learning and AI**: Bayes' theorem forms the basis of Bayesian inference, which is widely used in machine learning and artificial intelligence. It helps in updating hypotheses based on observed evidence, making it a key tool in probabilistic reasoning and decision-making under uncertainty.
3. **Probabilistic Models**: Bayes' theorem is essential in Bayesian statistics for building probabilistic models, where prior beliefs are updated with observed data to obtain posterior beliefs.
4. **What is a random variable, and how is it different from a regular variable?**

A random variable is a variable in probability theory and statistics that takes on numerical values as outcomes of a random phenomenon. It represents the numerical result of a random experiment or process, and its value is determined by chance.

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

The law of large numbers is a fundamental theorem in probability theory that describes the long-term behaviour of the average of a sequence of independent and identically distributed (i.i.d.) random variables. It states that as the number of trials (or observations) increases, the sample mean (average) of the observations converges to the expected value of the random variable.

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

Percentiles and quartiles are statistical measures used to summarize the distribution of a dataset by dividing it into equal parts or segments. They provide valuable insights into the spread and variation of data, especially in understanding the range of values and identifying potential outliers.

**Purpose of Using Percentiles and Quartiles:**

1. **Understanding Distribution**:
   * **Percentiles**: Percentiles divide a dataset into 100 equal parts. For instance, the 25th percentile (P25) represents the value below which 25% of the data falls. Percentiles help in understanding how data is spread across the entire range.
   * **Quartiles**: Quartiles divide a dataset into four equal parts:
     + First quartile (Q1) or 25th percentile,
     + Second quartile (Q2) or 50th percentile (which is also the median),
     + Third quartile (Q3) or 75th percentile.
   * Quartiles are particularly useful in summarizing the spread of data around the median and are less sensitive to extreme values compared to the mean and standard deviation.
2. **How do you use the central limit theorem to approximate a discrete probability distribution?**

The Central Limit Theorem (CLT) applies primarily to the sampling distribution of the sample mean from a population, especially when the sample size is large. It states that regardless of the population distribution, the sampling distribution of the sample mean tends to follow a normal distribution as the sample size increases. However, the CLT itself does not directly apply to approximating a discrete probability distribution in the traditional sense of discrete random variables.

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

Steps to Test Goodness of Fit:

**1. Specify the Hypothesized Distribution**:

**2**. **Formulate Hypotheses**:

**3**. **Collect and Organize Data**:

**4**. **Calculate Expected Frequencies or Probabilities**:

**5**. **Choose a Test Statistic**:

**6**. **Calculate the Test Statistic**:

**7**. **Determine the Critical Value or p-value**:

**8**. **Make a Decision**:

**28. What is a joint probability distribution?**

 **Joint Probability Distribution**: It describes the probabilities of multiple random variables occurring simultaneously.

 **Discrete Variables**: Described by joint probability functions pX,Y​(x,y) where X and Y are discrete.

 **Continuous Variables**: Described by joint probability density functions fX,Y​(x,y) where Xand Y are continuous.

 **Properties**: Includes marginal distributions and allows calculation of conditional probabilities and densities.

 **Application**: Essential for modelling dependencies between random variables and for statistical inference.

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

 **Joint Probability Distribution:** It specifies the probability of each possible combination of values for two or more random variables.

 **Marginal Probability Distribution:** It gives the probabilities of individual random variables without considering the values of the other variables.

**30. What is the difference between parameter estimation and hypothesis testing?**

**Parameter Estimation**:

* **Purpose**: Estimates unknown parameters (e.g., population mean, proportion) using sample data.
* **Outcome**: Provides a single value or interval estimate of the parameter.
* **Examples**: Calculating sample mean, confidence intervals.

**Hypothesis Testing**:

* **Purpose**: Tests a specific claim or hypothesis about a population parameter using sample data.
* **Outcome**: Determines if there is enough evidence to reject or fail to reject the null hypothesis.
* **Examples**: Comparing means, testing proportions, ANOVA tests.

Top of Form

Bottom of Form

**31. What is the p-value in hypothesis testing?**

**P-value in Hypothesis Testing**:

* **Definition**: Probability of obtaining results as extreme as the observed data, assuming the null hypothesis is true.
* **Interpretation**:
  + If p-value is low (e.g., less than 0.05), it suggests strong evidence against the null hypothesis, so it is rejected.
  + If p-value is high (e.g., greater than 0.05), it suggests weak evidence against the null hypothesis, so it is not rejected.
* **Conclusion**: Lower p-values indicate stronger evidence against the null hypothesis, supporting the alternative hypothesis.

**32. What is confidence interval estimation?**

Confidence interval estimation is a range of values around a sample statistic (like a mean or proportion) that likely includes the population parameter, with a specified level of confidence (such as 95% or 99%).

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

Type I error occurs when a null hypothesis is rejected incorrectly, meaning you conclude there is evidence for a relationship when there isn't.

Type II error happens when a null hypothesis is not rejected when it should be, indicating a failure to detect a relationship that actually exists.

**34. What is the difference between correlation and causation?**

Correlation refers to a relationship or association between two variables, where changes in one variable tend to coincide with changes in the other. Causation, on the other hand, implies that one variable directly causes a change in another variable. Correlation does not imply causation, as a relationship between variables can be influenced by other factors or occur by chance.

Top of Form

Bottom of Form

**35. How is a confidence interval defined in statistics?**

A confidence interval in statistics is a range of values calculated from sample data that is likely to contain the true population parameter with a specified level of confidence (e.g., 95% confidence).

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

The confidence level in a confidence interval represents the probability that the interval contains the true population parameter, assuming the sampling and estimation process is repeated multiple times. For example, a 95% confidence level means that if we were to construct 100 different confidence intervals from 100 different samples, approximately 95 of those intervals would contain the true population parameter.

**37. What is hypothesis testing in statistics?**

Bottom of Form

**Hypothesis Testing**:

* **Definition**: Statistical method to evaluate if there is enough evidence in a sample to infer a conclusion about a population parameter.
* **Process**: Involves formulating a null hypothesis (no effect) and an alternative hypothesis (effect exists), collecting data, and using statistical tests to assess the strength of evidence against the null hypothesis.
* **Purpose**: Provides a structured way to make decisions based on data, helping researchers draw conclusions about hypotheses.

Top of Form

Bottom of Form

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

**Purpose of Null Hypothesis**:

* **Definition**: States there is no significant difference or relationship between variables.
* **Role**: Serves as a baseline for comparison in hypothesis testing.
* **Testing**: Determines if the data provides enough evidence to reject the null hypothesis in favour of an alternative hypothesis.

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

**One-tailed test**:

* **Definition**: Tests for the possibility of an effect in one direction only (either positive or negative).
* **Application**: Used when there is a specific hypothesis about the direction of the effect.

**Two-tailed test**:

* **Definition**: Tests for the possibility of an effect in either direction (positive or negative).
* **Application**: Used when there is no specific hypothesis about the direction of the effect or when testing for differences in both directions.

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

**Experiment Design**:

* **Definition**: The process of planning and organizing an experiment to reliably test a hypothesis or research question.
* **Importance**:
  + Ensures validity: Proper design reduces bias and ensures results accurately reflect reality.
  + Enhances reliability: Consistent procedures and controls increase the likelihood of reproducible findings.
  + Maximizes efficiency: Clear design minimizes resource wastage and time spent on ineffective approaches.
  + Facilitates interpretation: Well-designed experiments yield clear, interpretable results that inform conclusions and future research directions.

Top of Form

Bottom of Form

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

**Key Elements in Experiment Design**:

* **Research question**: Clearly define what you want to investigate.
* **Variables**: Identify independent, dependent, and control variables.
* **Sampling**: Determine how participants or samples will be selected.
* **Randomization**: Randomly assign participants to groups to minimize bias.
* **Controls**: Implement measures to minimize confounding variables.
* **Replication**: Ensure the experiment can be repeated to verify results.
* **Ethics**: Consider ethical guidelines and participant rights.
* **Analysis**: Plan how data will be collected and analyzed.

**42. How can sample size determination affect experiment design?**

**Sample Size Determination and Experiment Design**:

* **Effect**: Influences the reliability and validity of study findings.
* **Key Points**:
  + Larger sample sizes generally improve the precision of estimates and increase statistical power.
  + Smaller sample sizes may lead to higher variability and less confidence in results.
  + Sample size calculations ensure studies have sufficient statistical power to detect meaningful effects.
  + Ethical considerations: Larger samples might involve more resources and participant burden.
  + Balancing: Determining an appropriate sample size involves trade-offs between precision, resources, and ethical considerations.

**43. What are observational and experimental data in statistics?**

**Observational Data**:

* **Definition**: Data collected by observing and recording events or phenomena as they naturally occur, without intervening or manipulating variables.
* **Example**: Surveys, census data, observational studies in social sciences.

**Experimental Data**:

* **Definition**: Data collected by intentionally manipulating one or more variables to observe their effect on other variables, often in controlled settings.
* **Example**: Clinical trials, randomized controlled trials (RCTs), laboratory experiments in sciences.

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

**Left-skewed distribution**:

* **Description**: Also called negatively skewed.
* **Characteristics**: The tail of the distribution extends to the left, with more data points on the right side of the peak.
* **Example**: Distribution of income (many people have low to moderate income, few have very high income).

**Right-skewed distribution**:

* **Description**: Also called positively skewed.
* **Characteristics**: The tail of the distribution extends to the right, with more data points on the left side of the peak.
* **Example**: Distribution of exam scores (many students score high, few score very low).

Top of Form

Bottom of Form

**45. What is Bessel’s correction?**

**Bessel's Correction**:

* **Purpose**: Adjusts the sample variance and sample standard deviation to provide an unbiased estimate of the population variance and population standard deviation.
* **Formula**: s2=∑(xi−xˉ)2/n−1 where n is the sample size.
* **Effect**: Without Bessel's correction, the sample variance tends to underestimate the population variance, especially for smaller sample sizes.

**46. What is kurtosis?**

**Kurtosis**:

* **Definition**: Measures the peakedness or flatness of a probability distribution compared to a normal distribution.
* **Positive Kurtosis**: Distribution is more peaked than normal (leptokurtic).
* **Negative Kurtosis**: Distribution is flatter than normal (platykurtic).
* **Zero Kurtosis**: Distribution has similar peakedness as a normal distribution (mesokurtic).

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

**Probability of throwing two fair dice**:

* **Sum is 5**: Probability = 4/36=1/9
  + Explanation: There are 4 combinations (1+4, 2+3, 3+2, 4+1) that sum to 5 out of 36 possible outcomes.
* **Sum is 8**: Probability = 5/36
  + Explanation: There are 5 combinations (2+6, 3+5, 4+4, 5+3, 6+2) that sum to 8 out of 36 possible outcomes.

**48. What is the difference between Descriptive and Inferential Statistics?**

**Descriptive Statistics**:

* **Purpose**: Summarizes and describes data.
* **Examples**: Mean, median, mode, range, standard deviation.
* **Use**: Provides insights into the characteristics of a dataset.

**Inferential Statistics**:

* **Purpose**: Makes inferences or predictions about a population based on sample data.
* **Examples**: t-tests, ANOVA, regression analysis.
* **Use**: Generalizes findings from a sample to a larger population.

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

**Degrees of Freedom (DF)**:

* **Meaning**: Represents the number of independent values or quantities which can be assigned to a statistical distribution or model, often related to the number of observations minus the number of constraints or parameters.
* **Application**: Essential in determining critical values for statistical tests like t-tests and chi-square tests, and in estimating variability in regression models.

**50. What is the empirical rule in Statistics?**

**Empirical Rule**:

* **Description**: Also known as the 68-95-99.7 rule.
* **Purpose**: Describes the approximate percentage of data within a certain number of standard deviations from the mean in a normal distribution.
* **Key Points**:
  + About 68% of the data falls within one standard deviation of the mean.
  + About 95% falls within two standard deviations.
  + About 99.7% falls within three standard deviations.

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

**Chi-Square Test**:

* **Purpose**: Tests if there is a significant association between categorical variables.
* **Types**: Goodness-of-fit (tests if observed frequencies match expected frequencies) and test of independence (tests if variables are independent).
* **Assumptions**: Expected frequencies in each category should be at least 5.
* **Output**: Chi-square statistic and p-value determine if the observed frequencies differ significantly from expected.
* **Application**: Used to analyse categorical data, such as survey responses or outcomes in contingency tables.

**52. What is a t-test?**

**T-test**:

* **Purpose**: Tests if the means of two groups are statistically different.
* **Types**: Independent t-test (between unrelated groups) and paired t-test (within related groups).
* **Assumptions**: Normally distributed data, similar variances between groups (for independent t-test).
* **Output**: t-statistic and p-value determine if group means differ significantly.
* **Application**: Commonly used in hypothesis testing to compare means in experiments or studies.

**53. What is the ANOVA test?**

**ANOVA (Analysis of Variance)**:

* **Purpose**: Tests if means of three or more groups are statistically different.
* **Types**: One-way (single factor) and two-way (two factors) ANOVA.
* **Assumptions**: Normal distribution within groups, homogeneity of variances, and independence of observations.
* **Output**: F-statistic and p-value determine if group means differ significantly.
* **Post hoc**: Follow-up tests (like Tukey's HSD) identify specific group differences after significant ANOVA results.

**54. How is hypothesis testing utilised in A/B testing for marketing campaigns?**

In A/B testing for marketing campaigns, hypothesis testing is used to determine whether the difference in outcomes between two variants (A and B) is statistically significant or if it could be due to random chance. This involves setting up null and alternative hypotheses based on the desired metric (e.g., conversion rate), collecting data from both variants, and using statistical tests (like t-tests or z-tests) to assess whether the observed difference is unlikely to have occurred by random variation alone. If the difference is statistically significant, it suggests that one variant performs better than the other, informing decisions on which variant to adopt for broader implementation.

**55. What is the difference between one-tailed and two tailed t-tests?**

The main difference between one-tailed and two-tailed t-tests lies in the hypotheses they test:

1. **One-tailed t-test**: Tests the hypothesis concerning the direction of the difference between groups. It examines whether the sample mean is significantly greater than or less than a known or hypothesized population mean, but not both.
2. **Two-tailed t-test**: Tests the hypothesis concerning the difference between groups in either direction. It examines whether the sample mean is significantly different from a known or hypothesized population mean, without specifying a direction.

In essence, one-tailed tests are more focused on detecting an effect in a specific direction, whereas two-tailed tests are more general and detect whether an effect exists regardless of direction.

Top of Form

Bottom of Form

**56. What is an inlier?**

An inlier is a data point that lies within a cluster or group of data points, typically in the context of outlier detection or clustering algorithms. It is considered a normal or typical observation within a dataset.