STATISTICS

Descriptive Statistics (Ds)

1. **What is the difference between the mean and the median of a data set?**
   The mean is the arithmetic average of all the values in a data set, while the median is the middle value when the data set is ordered. The mean is sensitive to extreme values (outliers), whereas the median is more robust to them.

2. **Define the term "standard deviation" and explain its significance in describing data variability.**
   Standard deviation measures the average distance of each data point from the mean. It quantifies the spread of the data, with larger values indicating more variability and smaller values indicating less variability.

3. **What is the mode of a data set, and how is it different from the mean?**
   The mode is the value that appears most frequently in a data set. Unlike the mean, which is calculated as the sum of all values divided by the number of values, the mode is simply the most common value and does not require any calculation.

4. **Explain the concept of skewness in a data distribution.**
   Skewness refers to the asymmetry of a data distribution. A distribution is negatively skewed if it has a longer tail on the left side, positively skewed if it has a longer tail on the right side, and symmetrical if both tails are of equal length.

5. **What is the interquartile range (IQR), and how is it used to identify outliers?**
   The IQR is the range between the first quartile (Q1) and the third quartile (Q3) of a data set. It measures the spread of the middle 50% of the data. Outliers are often identified as data points that lie below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR.

6. **How is variance different from standard deviation?**
   Variance is the average of the squared differences from the mean, while standard deviation is the square root of the variance. Both measure data variability, but standard deviation is in the same units as the data, making it more interpretable.

7. **What does the coefficient of variation tell us about a data set?**
   The coefficient of variation (CV) is the ratio of the standard deviation to the mean, expressed as a percentage. It provides a standardized measure of dispersion, allowing for comparison of variability across data sets with different units or scales.

8. **Describe what a boxplot represents in descriptive statistics.**
   A boxplot, or box-and-whisker plot, visualizes the distribution of a data set. It displays the median, quartiles, and potential outliers, with the box representing the IQR and the whiskers extending to the smallest and largest values within 1.5 * IQR.

9. **How do you calculate the range of a data set, and what does it indicate?**
   The range is calculated as the difference between the maximum and minimum values in a data set. It indicates the total spread or extent of the data.

10. **What is the purpose of a histogram, and how does it differ from a bar chart?**
    A histogram is used to represent the distribution of continuous numerical data by grouping

values into bins, whereas a bar chart is used for categorical data, with each bar representing a distinct category.

Inferential Statistics (is)

1. **What is hypothesis testing, and why is it important in statistical analysis?**
   Hypothesis testing is a method used to assess the evidence against a null hypothesis by analyzing sample data. It is important because it allows researchers to make inferences about a population based on sample data.
2. **Explain the concept of a confidence interval and how it is used to make inferences about a population.**
   A confidence interval is a range of values, derived from sample data, that is likely to contain the population parameter with a certain level of confidence (e.g., 95%). It provides an estimate of the parameter's uncertainty.
3. **What is the difference between a Type I and a Type II error in hypothesis testing?**
   A Type I error occurs when the null hypothesis is incorrectly rejected (false positive), while a Type II error occurs when the null hypothesis is incorrectly accepted (false negative).
4. **What does a p-value represent in the context of hypothesis testing?**
   A p-value represents the probability of observing the test statistic or more extreme values under the null hypothesis. A small p-value (typically < 0.05) suggests strong evidence against the null hypothesis.
5. **How do you interpret the results of a t-test?**
   A t-test compares the means of two groups to determine if they are significantly different. If the p-value is less than the chosen significance level (e.g., 0.05), we reject the null hypothesis and conclude that the means are significantly different.
6. **What is the central limit theorem, and why is it important in inferential statistics?**
   The central limit theorem states that the distribution of the sample mean will approach a normal distribution as the sample size increases, regardless of the population's distribution. This is important because it allows us to use normal distribution-based methods for inference.
7. **Describe the difference between a one-tailed and a two-tailed test in hypothesis testing.**
   A one-tailed test assesses the significance of deviations in one direction (either above or below the mean), while a two-tailed test assesses deviations in both directions. The choice depends on the research question.
8. **What is the purpose of an ANOVA test?**
   ANOVA (Analysis of Variance) is used to compare the means of three or more groups to determine if at least one group mean is significantly different from the others.
9. **Explain the concept of statistical power and how it affects hypothesis testing.**
   Statistical power is the probability that a test correctly rejects the null hypothesis when it is false (i.e., detects a true effect). Higher power reduces the risk of a Type II error.

10. **What is the difference between parametric and non-parametric tests?**
    Parametric tests assume underlying statistical distributions (e.g., normality), while non-parametric tests do not make such assumptions and are used when data do not meet these assumptions.

Probability Theory (pt)

1. **What is the difference between independent and dependent events in probability?**
   Independent events are those where the occurrence of one event does not affect the probability of the other event. Dependent events are those where the occurrence of one event affects the probability of the other.

2. **Define the term "expected value" and explain how it is calculated for a discrete random variable.**
   The expected value of a discrete random variable is the sum of all possible values weighted by their probabilities. It represents the long-term average outcome of a random process.

3. **What is the law of large numbers in probability theory?**
   The law of large numbers states that as the number of trials increases, the sample mean will converge to the expected value. This principle underpins the reliability of probability-based predictions in large samples.

4. **How is conditional probability different from joint probability?**
   Conditional probability is the probability of an event occurring given that another event has already occurred, while joint probability is the probability of two events occurring together.

5. **Explain the concept of mutually exclusive events in probability.**
   Mutually exclusive events are events that cannot occur simultaneously. If one event occurs, the other cannot, and the probability of both occurring together is zero.

6. **What is Bayes' theorem, and how is it used in probability?**
   Bayes' theorem describes how to update the probability of a hypothesis based on new evidence. It calculates the posterior probability of an event given prior knowledge and the likelihood of the evidence.

7. **Describe the difference between a probability distribution and a probability density function.**
   A probability distribution lists the probabilities of discrete outcomes, while a probability density function (PDF) describes the likelihood of continuous outcomes over a range.

8. **What is the difference between a discrete and a continuous random variable?**
   A discrete random variable takes on a countable number of distinct values, while a continuous random variable can take on an infinite number of values within a given range.

9. **How do you calculate the probability of the union of two events?**
   The probability of the union of two events is calculated as the sum of their individual probabilities minus the probability of their intersection: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

10. **What is the binomial distribution, and when is it used?**
    The binomial distribution describes the number of successes in a fixed number of

independent Bernoulli trials, each with the same probability of success. It is used when modeling binary outcomes (e.g., success/failure) over a fixed number of trials.