STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.
   a) True
   b) False
   Answer: a) True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
   a) Central Limit Theorem
   b) Central Mean Theorem
   c) Centroid Limit Theorem
   d) All of the mentioned
   Answer: a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
   a) Modeling event/time data
   b) Modeling bounded count data
   c) Modeling contingency tables
   d) All of the mentioned
   Answer: b) Modeling bounded count data

4. Point out the correct statement.
   a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
   b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   c) The square of a standard normal random variable follows what is called chi-squared distribution
   d) All of the mentioned
   Answer: d) All of the mentioned

5. _____random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
   Answer: c) Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
   Answer: b) False. Usually replacing the standard error by its estimated value doesn't change the CLT.

7. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned
   Answer: b) Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
   a) 0
   b) 5
   c) 1
   d) 10
   Answer: a) 0

9. Which of the following statement is incorrect with respect to outliers?
   a) Outliers can have varying degrees of influence
   b) Outliers can be the result of spurious or real processes
   c) Outliers cannot conform to the regression relationship
   d) None of the mentioned

   Answer: c) Outliers cannot conform to the regression relationship

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?

Answer: An example of a continuous probability distribution is the normal distribution, in which the majority of data points cluster around the middle of the range while the remaining ones taper off symmetrically towards either extreme. The mean distribution's is another name for the center of the range.

A Gaussian distribution or probability bell curve are other names for the normal distribution. Because it is symmetric around the mean, it shows that values close to the mean happen more frequently than those distant from the mean. A bell curve is what a normal distribution looks like graphically because of its flared shape. Depending on how values are distributed within the population, the exact shape may change. The entirety of the data points that make up the distribution is referred to as the population.

A normal distribution bell curve is always symmetrical around the mean, regardless of its precise shape. A symmetrical distribution has two mirror images on either side of a vertical dividing line that is drawn across the maximum/mean value, with half of the population being less than the mean and half being more. The converse, i.e., that all symmetrical distributions are normal, is not necessarily true. The mean, mode, and median of a bell curve are all the same, and the peak is always in the middle.

11. How do you handle missing data? What imputation techniques do you recommend?

Answer: There are many approaches to handle missing data. I think the most typical response is to disregard it. On the other hand, choosing to make no decision means that our statistical program will decide for us. Most of the time, our program will remove items in a listwise order. Listwise deletion may or may not be a wise choice, depending on why and how much data was lost.

Imputation is another tactic used frequently by people. Imputation involves replacing missing values with an estimate and assessing the complete set of data as if the imputed values were the actual observed values.

We can arrive at an estimate using the most popular techniques which include some of the following:

Mean imputation:

Determine the mean of all non-missing individuals' observed values for that variable. It has the benefit of keeping the mean and sample size constant, but it also has a number of disadvantages. The majority of the techniques listed below outperforms mean imputation.

Substitution

Assume the value comes from a brand-new individual who wasn't a part of the sample. Or, to put it another way, choose a different topic and use its merits.

Hot deck imputation

A value chosen at random from a sample participant who also has values for similar variables. To put it another way, identify all of the sample members who are comparable on other grounds, and then pick a random value from among their missing variable values.

The fact that you are restricted to only plausible values is one advantage. In other words, if the age range in your research is only allowed to be between 5 and 10, you will always get a value in this range. The random component, which adds some variation, introduces another element. This is essential for precise standard errors.

Cold deck imputation

A value that was purposefully chosen from someone who had comparable values for other factors. This is similar to Hot Deck in most ways, but without the random variance. You can always choose the third individual, for instance, using the same experimental setting and block.

Regression imputation

An anticipated value obtained by regressing the missing variable on other variables. As a result, you are relying on the predicted value, which is influenced by other factors, rather than using the mean. As a result, the imputation model's relationships between the variables are preserved, but not the variability around the predicted values.

Stochastic regression imputation

A regression's predicted value plus a random residual value. This combines the advantages of the

random component with those of regression imputation. Stochastic regression imputation is the foundation for the bulk of multiple imputation.

Interpolation and extrapolation

A judgement made on the basis of additional observations made by the same person. It typically only functions with data that has been gathered over time. But proceed with caution. Interpolation would make more sense for a variable like height in children—one that cannot be reduced through time. Extrapolation requires making more assumptions than necessary because it involves estimating outside the true range of the data.

Single or Multiple Imputation

- The two types of imputation are single and multiple. Imputation is typically used to refer to a single.
- The use of only one of the seven approaches to estimate the missing number described above is referred to as "single" estimation.
- It is well-liked since it is straightforward to comprehend and produces a sample with the same number of observations as the entire data set.
- Single imputation seems like a tempting alternative when listwise deletion removes a significant portion of the data set. It does, however, have some limitations.
- Certain imputation procedures, such as means, correlations, and regression coefficients, provide skewed parameter estimates unless the data is completely missing at random. The bias is frequently worse than when using listwise deletion, which is the default in the majority of software.
- The imputation method, the missing data mechanism, the percentage of missing data, and the information in the data set are some of the variables that affect the bias level.

Furthermore, all single imputation methods undervalue standard errors. The values of the imputed observations have a random error because they are estimates. When you enter that estimate as a data point, your program is oblivious of this, though. It overlooks the additional source of error as a result, producing p-values and standard errors that are excessively little.

Imputation is also challenging to grasp in practice despite being simple in theory. It isn't perfect as a result, but in some cases it may be adequate. Multiple imputation produces a large number of estimates. Two of the ways suggested using the above-hot deck and stochastic regression-work as the imputation method in multiple imputation.

Due to the inclusion of a random element in these two procedures, the multiple estimates varied greatly. As a result, some variance is once again introduced, which our program can take into consideration to produce accurate standard error estimates for our model.

Multiple imputation was a significant development in statistics about 20 years ago. When done effectively, it results in unbiased parameter estimations and precise standard errors while removing many (but not all) problems with missing data.

12. What is A/B testing?

Answer: A/B testing, also referred to as split testing, is a randomized experimentation process in which two or more variations of a variable (web page, page element, etc.) are displayed to various groups of website visitors at the same time to see which version has the greatest impact and influences business metrics.

A/B testing essentially takes all the guesswork out of website optimization and gives experience optimizers the ability to make data-backed judgements. In A/B testing, "control" or the original testing variable is referred to as "A." The term "variation" of the original testing variable is referred as B.

The "winner" is the version that causes your company metric(s) to change for the better. Your website can be optimized and your business ROI can be raised by implementing the changes of this winning variant on the page(s) or element(s) that we have already tested.

Each website has its own conversion stats. For instance, it may be the product sales in the context of eCommerce. In the meanwhile, it can be the creation of qualified leads for B2B.

One of the main steps in the Conversion Rate Optimization (CRO) process, A/B testing allows us to collect both qualitative and quantitative user data. With the help of the collected data, we may learn more about user behavior, engagement levels, problems, and even user satisfaction with new and improved website features. One will undoubtedly be losing out on a lot of potential business money if

they aren't A/B testing their website.

13. Is mean imputation of missing data acceptable practice?

Answer: Mean imputation is the process of replacing null values in a data collection with the mean of the data. Mean imputation is frequently seen as a bad practice since it disregards feature correlation. If we think about the following situation: we have a table containing age and fitness scores, but the fitness score for an eight-year-old is missing. The elderly person would appear to be far more fit than he actually is if we average the fitness scores of those between the ages of 15 and 80.

Second, mean imputation increases bias while reducing the variance of our data. The model is less accurate and the confidence interval is smaller as a result of the lower variance.

14. What is linear regression in statistics?

Answer: A variable's value can be predicted using linear regression analysis based on the value of another variable. The dependent variable is the one we want to be able to forecast. The independent variable is the one we're using to make a prediction about the value of the other variable.

With the help of one or more independent variables that can most accurately predict the value of the dependent variable, this type of analysis calculates the coefficients of the linear equation. The differences between expected and actual output values are minimized using linear regression by fitting a straight line or surface. The best-fit line for a set of paired data can be found using straightforward linear regression calculators that employ the "least squares" technique. The value of X (the dependent variable) is then estimated using Y (the independent variable).

15. What are the various branches of statistics?

Answer: The subject of statistics is divided into two subfields: descriptive statistics and inferential statistics.

Descriptive Statistics

Concept: This area of statistics is concerned with gathering, summarizing, and displaying data.

Example: The range in weight of 100 boxes of cereal randomly chosen from a factory's manufacturing line, the average age of voters who cast ballots for the winning candidate in the most recent presidential election, and the average length of all statistics books.

Interpretation: As there are so many instances in daily life, it is most probable that we are already familiar with this area of statistics. In a variety of disciplines, including securities trading, the social sciences, politics, the health sciences, and professional sports, descriptive statistics serve as the foundation for research and discussion. Using this area of statistics might frequently appear to be deceptively simple due to broad familiarity and availability of descriptive methods in many calculators and commercial applications. In descriptive statistics, the characteristics of sample and population data are described. The terms mean (average), variance, skewness, and kurtosis are used to describe statistical data.

Inferential Statistics

Concept: This area of statistics that examines sample information to make judgements about a population.

Example: The American Association of Retired Persons (AARP) conducted a survey of 2,001 full- or part-time employees aged 50 to 70, and found that 70% of respondents intended to continue working over the customary retirement age of mid-sixties. This figure could be used to make inferences about the population of all employees aged 50 to 70.

Interpretation: Whenever we use inferential statistics, you begin with a hypothesis and check to see if the data support it. Many inferential statistical techniques necessitate the use of a calculator or computer, and they can be readily misapplied or misinterpreted. Inferential statistics makes use of these characteristics to evaluate hypotheses and reach judgements. Analysis of variance (ANOVA),

logit/Probit models, and null hypothesis testing are examples of inferential statistics.