STATISTICS WORKSHEET-1

Q1. Bernoulli random variables take (only) the values 1 and 0.

Ans a) True

Q2. 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?

Ans a) Central Limit Theorem

Q3. Which of the following is incorrect with respect to use of Poisson distribution?

Ans b) Modeling bounded count data

Q4. 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

Ans d) All of the mentioned

Q5. _____ random variables are used to model rates.

Ans c) Poisson

Q6. Usually replacing the standard error by its estimated value does change the CLT.

Ans b) False

Q7. Which of the following testing is concerned with making decisions using data?

Ans b) Hypothesis

Q8. Normalized data are centered at_____and have units equal to standard deviations of the original data.

Ans a) 0

Q9. Which of the following statement is incorrect with respect to outliers?

Ans c) Outliers cannot conform to the regression relationship

Q10. What do you understand by the term Normal Distribution?

Ans

A normal distribution is the continuous probability distribution with a probability density function that gives you a symmetrical bell curve. Simply put, it is a plot of the probability function of a variable that has maximum data concentrated around one point and a few points taper off symmetrically towards two opposite ends.
In this definition of a normal distribution, you will explore the following terms:

- Continuous Probability Distribution: A probability distribution where the random variable, X, can take any given value, e.g., amount of rainfall. You can record the rainfall received at a certain time as 9 inches. But this is not an exact value. The actual value can be 9.001234 inches or an infinite amount of other numbers. There is no definitive way to plot a point in this case, and instead, you use a continuous value.

- Probability Density Function: An expression that is used to define the range of values that a continuous random variable can take.

A normal distribution has a probability distribution that is centered around the mean. This means that the distribution has more data around the mean. The data distribution decreases as you move away from the center. The resulting curve is symmetrical about the mean and forms a bell-shaped distribution.

Q11. How do you handle missing data? What imputation techniques do you recommend?

Ans

Missing data is an inevitable part of the process. As data researchers, we pour a lot of resources, time and energy into making sure the data set is as accurate as possible.

There are a lot of techniques to treat missing value.
Ignore the records with missing values.
Many tools ignore records with missing values. When the percentage of records with missing values is small, we could ignore those records.

- Substitute a value such as mean.

When the percentage is large and also when it makes sense to do something to avoid bias modeling results, substituting a value (e.g., mean, median) is a commonly used way. But this method could cause bias distribution and variance. That's where the following imputation methods come in.

- Predict missing values.

Depending on the type of the imputed variable (i.e., continuous, ordinal, nominal) and missing data pattern (i.e., monotone, non-monotone), below are a few commonly used models. If you plan to do it in SAS, there are SAS codes that you can write to identify the missing data pattern.

    - Logistic Regression
    - Discriminant Regression

- Predict missing values - Multiple Imputation. Although there are pros & cons, MI is considered to be superior to single imputation, and it better measures the uncertainty of the missing values.

In addition, there are a few required statistical assumptions for multiple imputation:

1. Whether the data is missing at random (MAR).
2. Multivariate normal distribution, for some of the modeling methods mentioned above (e.g., regression, MCMC).
3. The type of imputation algorithm used.

- Some justification for choosing a particular imputation method.
- The proportion of missing observations.
- The number of imputed datasets (m) created.
- The variables used in the imputation model.

Q12. What is A/B testing?

Ans

A/B testing in its simplest sense is an experiment on two variants to see which performs better based on a given metric. Typically, two consumer groups are exposed to two different versions of the same thing to see if there is a significant difference in metrics like sessions, click-through rate, and/or conversions.

If, we could randomly split our customer base into two groups, a control group and a variant group. Then, we can expose our variant group with a red website banner and see if we get a significant increase in conversions. It's important to note that all other variables need to be held constant when performing an A/B test.

Getting more technical, A/B testing is a form of statistical and two-sample hypothesis testing. Statistical hypothesis testing is a method in which a sample dataset is compared against the population data. Two-sample hypothesis testing is a method in determining whether the differences between the two samples are statistically significant or not.

Q13. Is mean imputation of missing data acceptable practice?

Ans

The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation.
Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

Q14. What is linear regression in statistics?

Ans

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

Linear regression is a kind of statistical analysis that attempts to show a relationship between two variables. Linear regression looks at various data points and plots a trend line. Linear regression can create a predictive model on apparently random data, showing trends in data, such as in cancer diagnoses or in stock prices.

Linear regression is an important tool in analytics. The technique uses statistical calculations to plot a trend line in a set of data points. The trend line could be anything from the number of people diagnosed with skin cancer to the financial performance of a company.
Linear regression shows a relationship between an independent variable and a dependent variable being studied.

There are a number of ways to calculate linear regression. One of the most common is the ordinary least-squares method, which estimates unknown variables in the data, which visually turns into the sum of the vertical distances between the data points and the trend line.
The calculations to perform linear regressions can be quite complex. Fortunately, linear regression models are included in most major calculation’s packages, such as Excel, R, MATLAB and Mathematica.

Q15. What are the various branches of statistics?

Ans

Statistics is a study of presentation, analysis, collection, interpretation and organization of data

There are two main branches of statistics
- Inferential Statistic.
- Descriptive Statistic.

Inferential Statistics:
Inferential statistics used to make inference and describe about the population. These stats are more useful when it's not easy or possible to examine each member of the population.

Descriptive Statistics:
Descriptive statistics are used to get a brief summary of data. You can have the summary of data in numerical or graphical form.