STATISTICS WORKSHEET-1

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

Answer: a) True

b) False

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?

Answer: a) Central Limit Theorem

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

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

a) Modelling event/time data

Answer: b) Modelling bounded count data

c) Modelling contingency tables

d) All of the mentioned

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

Answer: d) All of the mentioned

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

Answer: c) Poisson

d) All of the mentioned

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

a) True

Answer: b) False

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

a) Probability

Answer: b) Hypothesis

c) Causal

d) None of the mentioned

8. Normalized data are centred at _____ and have units equal to standard deviations of the original data.

Answer: a) 0

b) 5

c) 1

d) 10

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

Answer: c) Outliers cannot conform to the regression relationship

d) None of the mentioned

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

Answer: The normal distribution, also known as Gaussian distribution, is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. A graphical representation of a normal distribution is sometimes called a bell curve because of its flared shape.

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

Answer: When dealing with missing data, data scientists can use two primary methods to solve the error: imputation or the removal of data.

The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

Imputation techniques:

1. ***Mean, Median and Mode***

This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, data scientists can calculate the mean or median of the existing observations.

2. ***Time-Series Specific Methods***

Another option is to use time-series specific methods when appropriate to impute data. There are four types of time-series data:

- No trend or seasonality.
- Trend, but no seasonality.
- Seasonality, but no trend.
- Both trend and seasonality.

3. ***Last Observation Carried Forward (LOCF) & Next Observation Carried Backward (NOCB)***

These options are used to analyse longitudinal repeated measures data, in which follow-up observations may be missing. In this method, every missing value is replaced with the last observed value. Longitudinal data track the same instance at different points along a timeline.

4. ***Linear Interpolation***

Linear interpolation is often used to approximate a value of some function by using two known values of that function at other points. This formula can also be understood as a weighted average. The weights are inversely related to the distance from the end points to the unknown point. The closer point has more influence than the farther point.

5. *Seasonal Adjustment with Linear Interpolation*

When dealing with data that exhibits both trend and seasonality characteristics, we use seasonal adjustment with linear interpolation.

12. What is A/B testing?

Answer: A/B testing (also known as bucket testing or split-run testing) is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

13. Is mean imputation of missing data acceptable practice?

Answer: Mean imputation is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, data scientists can calculate the mean or median of the existing observations. However, when there are many missing variables, mean or median results can result in a loss of variation in the data. This method does not preserve the relationships among variables. It also leads to an underestimate of standard errors. Hence, mean imputation is generally not recommended.

14. What is linear regression in statistics?

Answer: Linear regression is a sub-category of regression (which deals with numerical data) in Statistics and it 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).

15. What are the various branches of statistics?

Answer: Descriptive statistics and inferential statistics are the two main branches of statistics. Both of these are used in scientific data analysis.

Descriptive statistics deals with the presentation and collecting of data. It is not so simple as it appears, and the statistician must be aware of how to design and experiment, select the appropriate focus group, and prevent biases that are all too easy to introduce into the experiment.

Generally, descriptive statistics can be categorized into:

- Central tendency (Mean, median, mode)
- Spread of data (Quartiles, ranges, variances, skew, and standard deviation)

Inference statistics are statistical techniques that allow statisticians to utilise data from a sample to conclude, predict the behaviour of a given population, and make judgments or decisions.

Using descriptive statistics, inference statistics frequently talk in terms of probability. Furthermore, a statistician uses these techniques mainly for data analysis, writing, and drawing conclusions from the limited data. This is accomplished by taking samples and determining their reliability.

Most future predictions and generalisations based on a population study of a smaller specimen are covered by inference statistics. Furthermore, the majority of social science experiments involve the investigation of a small sample population and that helps in determining community behaviour.

The researchers can bring the study related conclusions by designing a practical experiment. When drawing conclusions, it is important to avoid drawing incorrect or biased conclusions.

And there are some of the different types of inferential statistics which includes the following which are shown below:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Statistical significance (t-test)
- Correlation analysis