STATISTICS WORKSHEET-1

Answers of multiple choice:

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

3. Which of the following is incorrect with respect to use of Poisson distribution?
b) Modeling bounded count data

4. Point out the correct statement.
d) All of the mentioned

5. _____ random variables are used to model rates.
c) Poisson

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

7. Which of the following testing is concerned with making decisions using data?
b) Hypothesis
c) Causal

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
a) 0

9. Which of the following statement is incorrect with respect to outliers?
c) Outliers cannot conform to the regression relationship

Subjective answers-

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

Answer: A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

Height is one simple example of something that follows a normal distribution pattern: Most people are of average height, the numbers of people that are taller and shorter than average are fairly equal and a very small (and still roughly equivalent) number of people are either extremely tall or extremely short.

Here's an example of a normal distribution curve:
A graphical representation of a normal distribution is sometimes called a bell curve because of its flared shape. The precise shape can vary according to the distribution of the population but the peak is always in the middle and the curve is always symmetrical. In a normal distribution, the mean, mode and median are all the same. Normal distribution curves are sometimes designed with a histogram inside the curve. The graphs are commonly used in mathematics, statistics and corporate data analytics.

11. How do you handle missing data? What imputation techniques do you recommend?
Answer: The cause of missing values can be data corruption or failure to record data. Best techniques to handle missing data are:

- Deleting Rows with missing values
- Impute missing values for continuous variable
- Impute missing values for categorical variable
- Other Imputation Methods
- Using Algorithms that support missing values
- Prediction of missing values
- Imputation using Deep Learning Library- Datawig

Common Imputation techniques, which can be recommended, are as follows:

a) Mean or Median Imputation: When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons why this may not be the most feasible option:

- There may not be enough observations with non-missing data to produce a reliable analysis
- In predictive analytics, missing data can prevent the predictions for those observations which have missing data
- External factors may require specific observations to be part of the analysis

b) Multivariate Imputation by Chained Equations (MICE): MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.

c) Random Forest: Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.

12. What is A/B testing?

Answer: A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

A/B is testing is important because, it helps in following-

- Solve visitor pain points
- Get better ROI from existing traffic
- Reduce bounce rate
- Make low-risk modifications
- Achieve statistically significant improvements
- Redesign website to increase future business gains

13. Is mean imputation of missing data acceptable practice?

Answer: No mean imputation of missing data is not acceptable practice. Below are the three problems with using mean-imputed variables in statistical analyses:

- Mean imputation reduces the variance of the imputed variables.
- Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.
- Mean imputation does not preserve relationships between variables such as correlations.

14. What is linear regression in statistics?

Answer: Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

A linear regression line has an equation of the form $\boldsymbol{Y = a + bX}$, where $\boldsymbol{X}$ is the explanatory variable and $\boldsymbol{Y}$ is the dependent variable. The slope of the line is $\boldsymbol{b}$, and $\boldsymbol{a}$ is the intercept (the value of $\boldsymbol{y}$ when $\boldsymbol{x} = 0$).

Types of Linear Regression are-

a) **Simple linear regression**: 1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)
b) **Multiple linear regression**: 1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)
c) **Logistic regression**: 1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)
d) **Ordinal regression**: 1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)
e) **Multinomial regression**: 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

f) **Discriminant analysis**: 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

15. What are the various branches of statistics?

Answer: The two main branches of statistics are descriptive statistics and inferential statistics.

1. **Descriptive Statistics**: Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, a nd the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment. Descriptive statistics can be categorized into-

- Measures of central tendency
- Measures of variability

Measures of Central Tendency

Measures of central tendency specifically help the statisticians to estimate the center of values distribution. These measures of tendency are:

- **Mean**: This is the conventional method used in describing central tendency. Usually, to compute an average of values, you add up all the values and then divide them with the number of values available.
- **Median**: This is the score found at the middle of a set of values. A simple way to calculate a median is to arrange the scores in numerical orders and then locate the score, which is at the center of the arranged sample.
- **Mode**: This is the frequently occurring value in a given set of scores.

Measures of Variability

The measure of variability help statisticians to analyze the distribution spread out of a given set of data. Some of the examples of measures of variability include quartiles, range, variance and standard deviation.

2. **Inferential Statistics**: Inferential statistics are techniques that enable statisticians to use the gathered information from a sample to make inferences, decisions or predictions about a given population. Inferential statistics often talks in probability terms by using descriptive statistics. Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics.
   The different types of calculation of inferential statistics include:

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