STATISTICS WORKSHEET_1

1. A
2. A
3. B
4. D
5. C
6. B
7. B
8. A
9. C
10. In probability theory, a normal distribution is a type of continuous probability distribution for a real valued random variable. Normal distribution is sometimes called as bell curve. The general form for probability density function of normal distribution is:

$$f(x)=e^{-(x-\mu)2/(2\sigma 2)}\sigma\sqrt{2\pi}$$

11. Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. The common methods to deal with missing data are:
    1) Mean or Median Imputation
       A common technique is to use the mean or median of a particular column to replace missing data. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations
    2) Multivariate Imputaions 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 the missing data.
    3) 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 out of bag imputation error estimates.
12. A/B testing also known as bucket testing or split-run testing is a user experience research methodology. It consists of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or two-sample hypothesis testing used in the field of statistics. A/B testing is 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. Mean imputation of missing data is sometimes acceptable in practice but mostly to get accurate result it is not recommended because for large number of missing values, it can result in loss of variation in data. So for data with low number of missing values we could use the mean imputation of missing data.

14. 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).
15. Statistics is divided into two main branches:
    1) Descriptive Statistics
       It deals with collection of data and presentation of data in various forms such as, tables, graphs, and diagrams and finding averages and other measures which would describe data.
    2) Inferential Statistics
       It deals with techniques used for analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates.