**STATISTICS** WORKSHEET **1**

1)A-True

2)A- central limit theorem

3) D

4)D

5)C- Poisson

6)B- false

7)B- hypothesis

8)A-0

9) D- none of the mentioned

10) Normal Distribution: A Normal Distribution is also known as a Gaussian distribution or famously Bell Curve. The probability density function (pdf) for Normal Distribution:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi * \sigma^2}} * e^{-\frac{1}{2} * (\frac{x-\mu}{\sigma})^2}$$

- Symmetric distribution is also known as Normal distribution where Mean=Median=Mode.
- All the continuous data follow normal distribution.
- The normal distribution is a form presenting data by arranging the probability distribution of each value in the data. Most values remain around the mean value making the arrangement symmetric.

11) Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. Data can be missing in the following ways:

**Missing Completely At Random (MCAR):** When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random. A quick check for this is to compare two parts of data – one with missing observations and the other without missing observations. On a t-test, if we do not find any difference in means between the two samples of data, we can assume the data to be MCAR.

**Missing At Random (MAR):** The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data. For example, if high school GPA data is missing randomly across all schools in a district, that data will be considered MCAR. However, if data is randomly missing for students in specific schools of the district, then the data is MAR.

**Not Missing At Random (NMAR**): When the missing data has a structure to it, we cannot treat it as missing at random. In the above example, if the data was missing for all students from specific schools, then the data cannot be treated as MAR.

It is important to understand the nature of the data that is missing when deciding which algorithm to use for imputations. While using the above algorithms, predictor variables should be set up carefully to avoid confusion in the methods implemented during imputation. Finally, we can test the quality of your imputations by normalized root mean square error (NRMSE) for continuous variables and proportion of falsely classified (PFC) for categorical variables.

12) A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

13) Due to below problems mean imputation of missing data is not acceptable practice:

a) Mean Imputation Leads to An Underestimate of Standard Errors

b) Mean imputation does not preserve the relationships among variables

14) Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

15) There are two main branches of statistics: **Descriptive statistics and inferential statistics**.

**Descriptive Statistics**

CONCEPT The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

EXAMPLES The average age of citizens who voted for the winning candidate in the last presidential election, the average length of all books about statistics, the variation in the weight of 100 boxes of cereal selected from a factory's production line.

**Inferential Statistics**

CONCEPT The branch of statistics that analyzes sample data to draw conclusions about a population.

EXAMPLE A survey that sampled 2,001 full-or part-time workers ages 50 to 70, conducted by the American Association of Retired Persons (AARP), discovered that 70% of those polled planned to work past the traditional mid-60s retirement age. By using methods discussed in Section 6.4, this statistic could be used to draw conclusions about the population of all workers ages 50 to 70.