**STATISTICS WORKSHEET**

**Q1.** A

**Q2.** A

**Q3.** B

**Q4.** D

**Q5.** C

**Q6.** B

**Q7.** B

**Q8.** A

**Q9.** C

**Q10.** 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 number of people are either extremely tall or extremely short.

# Q11. Types of Missing Data

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.

**Common Methods**

**1. 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

In such cases, we impute values for missing data. A common technique is to use the mean or median of the non-missing observations. 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. Depending upon the nature of the missing data, we use different techniques to impute data that have been described below.

**2. 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 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.

To set up the data for MICE, it is important to note that the algorithm uses all the variables in the data for predictions. In this case, variables that may not be useful for predictions, like the ID variable, should be removed before implementing this algorithm.

**Q12.** 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.

**Example:** 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.

It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The**population** refers to all the customers buying your product, while the **sample** refers to the number of customers that participated in the test.

**Q13.** It is a non-standard, but a fairly flexible imputation algorithm. It uses RandomForest at its core to predict the missing data. It can be applied to both continuous and categorical variables which makes it advantageous over other imputation algorithms.

**Q14.** 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 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.

There are many names for a regression’s dependent variable.  It may be called an outcome variable, criterion variable, endogenous variable. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are

(1) determining the strength of predictors,

(2) forecasting an effect,

(3) trend forecasting.

**Q15.** The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

## Descriptive Statistics:

[Descriptive statistics](https://explorable.com/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, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid [biases](https://explorable.com/research-bias) that are so easy to creep into the [experiment](https://explorable.com/conducting-an-experiment).

Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

## Inferential Statistics:

[Inferential statistics](https://explorable.com/inferential-statistics), as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

Most [predictions](https://explorable.com/prediction-in-research) of the future and [generalizations](https://explorable.com/what-is-generalization) about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small [sample population](https://explorable.com/sample-group) that helps determine how the population in general behaves.

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good [scientific methodology](https://explorable.com/research-methodology) needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.