# Flip Robo

## Statistics Worksheet - 1 Solutions

**1.** Correct option - a & Correct Answer - True

**2.** Correct option - a & Correct Answer - Central limit theorem

**3.** Correct option - b & Correct Answer - Modeling bounded count data

**4.** Correct option - d & Correct Answer - All of the mentioned

**5.** Correct option - c & Correct Answer - poisson

**6.** Correct option - b & Correct Answer - False

**7.** Correct option - b & Correct Answer - Hypothesis

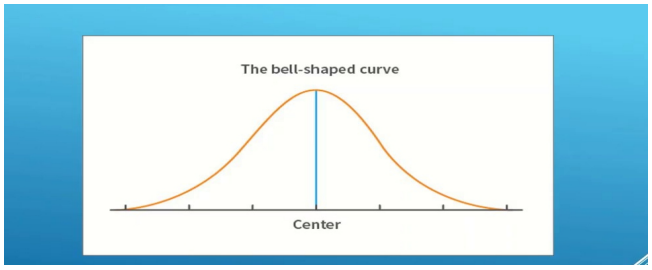
**8.** Correct option - a & Correct Answer - 0

**9.** Correct option - c & Correct Answer - outliers cannot conform to the regression relationship.

**10.** A probability function that specifies how the values of a variable are distributed is called the normal distribution. It is symmetric since most of the observations assemble around the central peak of the curve. The probabilities for values of the distribution are distant from the mean narrow off evenly in both directions. **Normal Distribution** also called the **Gaussian Distribution**, is the most significant continuous probability distribution. Sometimes it is also called a bell curve.
Some of the important properties of the normal distribution are listed below:

- In a normal distribution, the mean, mean and mode are equal.(i.e., Mean = Median= Mode).

- The total area under the curve should be equal to 1.
- The normally distributed curve should be symmetric at the centre.
- There should be exactly half of the values are to the right of the centre and exactly half of the values are to the left of the centre.
- The normal distribution should be defined by the mean and standard deviation.
- The normal distribution curve must have only one peak. (i.e., Unimodal)
- The curve approaches the x-axis, but it never touches, and it extends farther away from the mean.




**11.** Missing data can skew anything for data scientists, from economic analysis to clinical trials. After all, any analysis is only as good as the data. A data scientist doesn't want to produce biased estimates that lead to invalid results. The concept of missing data is implied in the name: it's data that is not captured for a variable for the observation in question. Missing data reduces the statistical power of the analysis, which can distort the validity of the results. Before jumping to the methods of data imputation, we have to understand the reason why data goes missing.

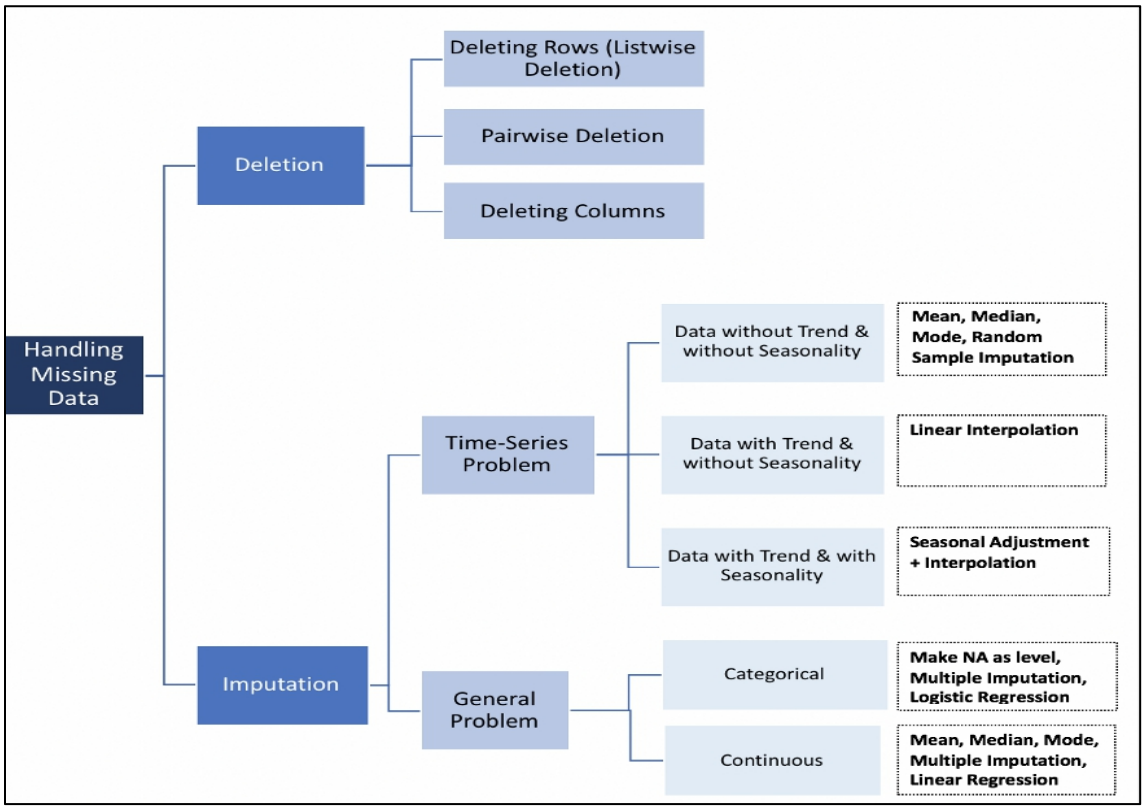
1. **Missing at Random (MAR):** Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data.

2. **Missing Completely at Random (MCAR):** The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.

3. **Missing not at Random (MNAR):** Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some

other variable's value (e.g. Let's assume that females generally don't want to reveal their ages! Here the missing value in age variable is impacted by gender variable).

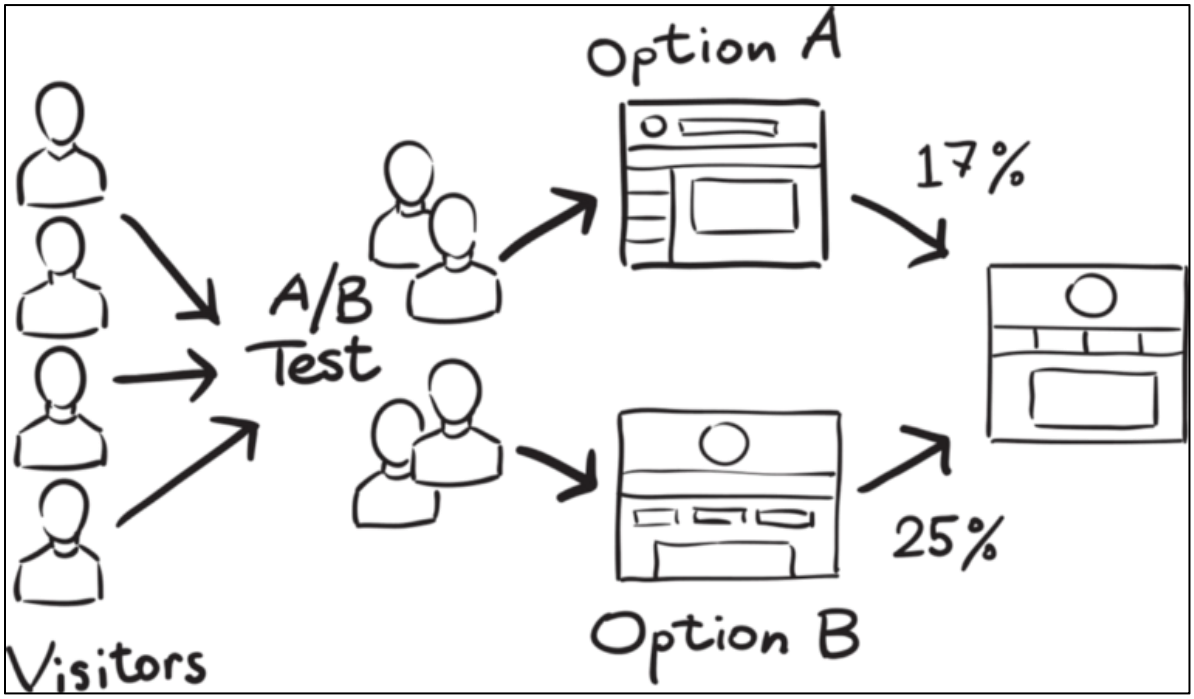
In the first two cases, it is safe to remove the data with missing values depending upon their occurrences, while in the third case removing observations with missing values can produce a bias in the model. So we have to be really careful before removing observations. Note that imputation does not necessarily give better results.




From the above figure, we can see the imputation techniques which are used to handle the missing data problems.

**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.** Mean imputation is the practice of replacing null values in a data set with the mean of the data.Mean imputation is generally bad practice because it doesn't take into account feature correlation. For example, imagine we have a table showing age and fitness score and imagine that an eighty-year-old has a missing fitness score. If we took the average fitness score from an age range of 15 to 80, then the eighty-year-old will appear to have a much higher fitness score that he actually should.Second, mean imputation reduces the variance of the data and increases bias in our data. This leads to a less accurate model and a narrower confidence interval due to a smaller variance.

**14.** Linear regression is a statistical technique that is used to learn more about the relationship between an independent (predictor) variable and a dependent (criterion) variable. When you have more than one independent variable in your analysis, this is referred to as multiple linear regression. For example, let say we were studying the causes of obesity, measured by body mass index (BMI). In particular, we wanted to

see if the following variables were significant predictors of a person's BMI: number of fast food meals eaten per week, number of hours of television watched per week, the number of minutes spent exercising per week, and parents' BMI. Linear regression would be a good methodology for this analysis.

## 15. Branches of Statistics

1. Descriptive Statistics
a. Measures of Central Tendency
b. Measures of Variability
2. Inferential Statistics

**Descriptive Statistics**

Descriptive statistics is the first part of statistics that deals with the collection of data. People seem it too easy, but it is not that easy. The statisticians need to be aware of the designing and experiments. They also need to choose the right focus group and avoid biases. In contrast, Descriptive statistics are used in use to do various kinds of analysis on different studies.

**Descriptive statistics have two parts**

- Central tendency measures
- Variability measures

To help understand the analyzed data, the tendency measures and variability measures use tables, general discussions, and charts.

Measures of Central Tendency

Central tendency measures specifically help statisticians evaluate the distribution center of values. These tendency measures are:

**Mean:-** Mean is a conventional method used to describe the central tendency. Typically, to calculate the average of values, count all values, and then divide them with the number of available values.

**Median:-** It is the result that is in the middle of a set of values. An easy way to calculate the median is to edit the results in numerical journals and locate the result that is in the center of the distributed sample.

**Mode:-** The mode is the frequently occurring value in the given data set.

**Measures of Variability**

The variability measure helps statisticians to analyze the distribution that is spreading from a specific data set. Some of the variables of variability include quartiles, ranges, variances, and standard deviation.

**Inferential Statistics**

The inference statistics are techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population.

Inference statistics often speak in terms of probability by using descriptive statistics. Besides, these techniques are used primarily by a statistician for data analysis, drafting, and making conclusions from limited information. That is obtained by taking samples and testing how reliable they are.

Most predictions of the future and generalization on a population study of a smaller specimen are in the scope of the inference statistics. Besides, most of the social sciences experiments deal with the study of a small sample population that helps determine the behavior of the community.

Designing a real experiment, the researcher can bring conclusions relevant to his study. When making conclusions, it should be cautious not to draw wrongly or biased

Different types of inferential statistics include:

- Regression analysis
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)

- Statistical significance (t-test)
- Correlation analysis