# Statistics Worksheet

**Q1.** Bernoulli random variables take (only) the values 1 and 0.

Sol: Option A i.e True

**Q2.** 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?

Sol: Option A i.e Central Limit Theorem

**Q3.** Which of the following is incorrect with respect to use of Poisson distribution?

Sol: Option B i.e Modeling bounded count data

**Q4**. Point out the correct statement.

Sol: Option D i.e All of the mentioned above

**Q5**. _____random variables are used to model rates

Sol: Option C i.e Poisson

**Q6**. Usually replacing the standard error by its estimated value does change the CLT.

Sol: Option B i.e False

**Q7**. Which of the following testing is concerned with making decisions using data?

Sol: Option B i.e Hypothesis

**Q8**. Normalized data are centered at_____and have units equal to standard deviations of the original data.

Sol: Option A i.e 0

**Q9**. Which of the following statement is incorrect with respect to outliers?

Sol: Option C i.e Outliers cannot conform to the regression relationship

**Q10. What do you understand by the term Normal Distribution?**

Sol: The Normal Distribution is defined by the Probability Density Function, which is, the density for a continuous random variable lying between a specific range of values in a system. It describes how the values of a variable are distributed. Normal distribution is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are unlikely.

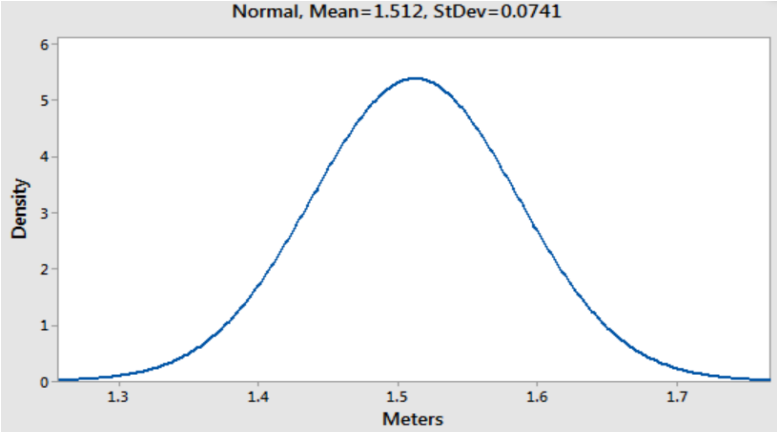
The normal distribution has two parameters, the mean and standard deviation.

**Mean:** The mean is the central tendency of the normal distribution. It defines the location of the peak for the bell curve. Most values cluster around the mean. Change in mean shifts the entire curve left or right on the X-axis.

**Standard deviation:** The standard deviation is a measure of variability. It defines the width of the normal distribution. The standard deviation determines how far away from the mean the values tend to fall. It represents the typical distance between the observations and the average. Change in standard deviation either tightens or spreads out the width of the distribution along the X-axis.

Example of Normally Distributed Data

Height data are normally distributed. Below is the data collected from few girls during a study.




- The distribution of heights follows the typical bell curve pattern for all normal distributions. Most girls are close to the average (1.512 meters).

- Small differences between an individual's height and the mean occur more frequently than substantial deviations from the mean. The standard deviation is 0.0741m, which indicates the typical distance that individual girls tend to fall from mean height.

- The distribution is symmetric. The number of girls shorter than average equals the number of girls taller than average. In both tails of the distribution, extremely short girls occur as infrequently as extremely tall girls.

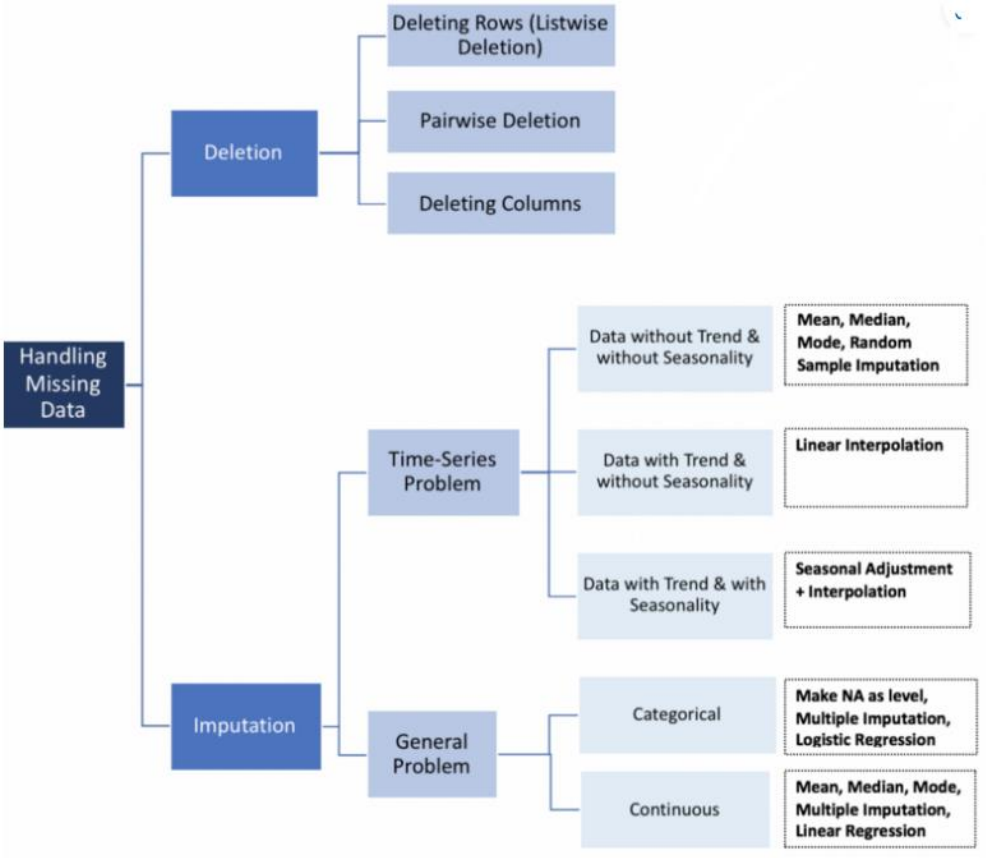
**Q11. How do you handle missing data? What imputation techniques do you recommend?**

Sol: Missing Data, or missing values, occur when the data is not present for certain variables. Data can go missing due to incomplete data entry, equipment malfunctions, lost files,etc. Missing data are errors as the data don't represent the true values of what it set out to measure.

Moreover, if the values are missing completely at random, the data sample is likely still representative of the population. But if the values are missing systematically, analysis may be biased.

| Type | Definition |
|---|---|
| Missing completely at random (MCAR) | Missing data are randomly distributed across the variable and unrelated to other variables. |
| Missing at random (MAR) | Missing data are not randomly distributed but they are accounted for by other observed variables. |
| Missing not at random (MNAR) | Missing data systematically differ from the observed values. |

- To handle the missing data, there is no particular way, that can be generalized. However, different solutions for data imputation depending on the kind of problem.




- The best imputation that can be recommended for missing data is the creation of Dummy Data. It is mock data generated randomly as a substitute for missing places that assists in removing biasness from any other method.

**Q12. What is A/B testing?**

Sol: A/B testing, also known as bucket testing or split-run testing is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a 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. For instance, the randomized experimentation process of testing wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

A/B testing is a shorthand for a simple randomized controlled experiment, in which two samples (A and B) of a single vector-variable are compared. These values are similar except for one variation which might affect a user's behaviour. A/B tests are widely considered the simplest form of controlled experiment. However, by adding more variants to the test, its complexity grows.

- A/B tests are useful for understanding user engagement and satisfaction of online features like a new feature or product.
- Large social media sites like LinkedIn, Facebook, and Instagram use A/B testing to make user experiences more successful and as a way to streamline their services.

**Q13.Is mean imputation of missing data acceptable practice?**

Sol: The Mean Imputation where the missing value is replaced for the mean of all data formed within a specific cell or class. This technique isn't a good idea because the mean is sensitive to data noise like outliers. Further,

- 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.

**Q14.What is linear regression in statistics?**

Sol: Linear regression is used to predict the relationship between two variables by applying a linear equation to observed data. There are two types of variable, one variable is called an independent variable, and the other is a dependent variable. Linear regression is commonly used for predictive analysis.

The main 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 are significant predictors of 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

$$y = c + b*x,$$

where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent 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).

**Q15.What are the various branches of statistics?**

Sol: **Statistics** is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data. In other words, statistics is a form of mathematical analysis that uses quantitative models to give a set of experimental data or . Statistics examine the methodology for collecting, reviewing, analyzing, and making data conclusions. Some statistical measures include the following:

- **Mean:** It is an important concept in mathematics and statistics. The mean is an average and the most common value in the collection of numbers.
- **Regression analysis:** It is a powerful statistical method. It allows us to examine the relationship between two or more variables of interest.
- **Skewness:** In statistics, skewness is a degree of asymmetry that is observed in a probability distribution. Distributions can display right (positive) skewness or left (negative) skewness to differing degrees. A normal distribution (bell curve) presents zero skewness.
- **Kurtosis**: It is a measure of the combined weight of a distribution's tails relative to the centre of the distribution.
    - Variance: It estimates the variability from the mean or average.
    - Analysis of variance: The method of statistics that separates the variance data into several components used for additional tests.

The two main branches of statistics are descriptive statistics and inferential statistics **.**

**Descriptive statistics** deals with the collection of data, its presentation in various forms, such as tables, graphs and diagrams and finding averages and other measures which would describe the data.

For example: Industrial statistics, population statistics, trade statistics, etc. Businessmen make use of descriptive statistics in presenting their annual reports, final accounts, and bank statements.

**Inferential statistics** deals with techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates. It makes inference and prediction about population based on a sample of data taken from population. It generalizes a large dataset and applies probabilities to draw a conclusion. It is simply used for explaining meaning of descriptive statistics.

For example: Suppose, to have an idea about the percentage of the illiterate population of the country. Taking a sample from the population and find the proportion of illiterate individuals

in the sample. With the help of probability, this sample proportion enables us to make some inferences about the population proportion. This study belongs to inferential statistics.