# STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question,

1. Bernoulli random variables take (only) the values 1 and O.

a) True
b) False

2.Which of the following theorem states that the distribution of averages of id variables, properly normalized, becomes that of a standard normal as the sample size increases?

a) Central Limit Theorem
b) Central Mean Theorem
c) Centroid Limit Theorem
d) All of the mentioned

3.Which of the following is incorreet with respeet to use of Poison distribution?

a) Modeling event time data
b) Modeling bounded count data
c) Modeling contingency tables
d) All of the mentioned

4.Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log- normal
distribution
b) Sums of normally distributed random variables are again normally distributed even if the variables
are dependent
c) The square of a standard normal random variable follows what is called chi-squared
distribution
d) All of the mentioned

5. Random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

6. Usually replacing the standard etror by its estimated value does change the CLT,
a) True
b) False

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

a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

8.Normalized data are centered at
and have units equal to standard deviations of the
original data.

a) 0
b) 5
c) I
d) 10

9.Which of the following statement is incorreet with respect to outliers?

a)Outliers can have varying degrees of influence
b)Outliers can be the result ot spurious or real processes
C)Outliers cannot conform to the regression relationship
d) None of the mentioned

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?
11. How do you handle missing data? What imputation-techniques do you recommend?
12What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14What is linear regression in statistics?
15. What are the various branches of statistics?

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

**Ans:-** Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

The normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses. The standard normal distribution has two parameters: the mean and the standard deviation. For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations.

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

**Ans:-** When dealing with missing data, we can use two primary methods to solve the error: imputation or the removal of data.

The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required.

Before deciding which approach to employ, data scientists must understand why the data is missing.

### Missing at Random (MAR)

Missing at Random means the data is missing relative to the observed data. It is not related to the specific missing values. The data is not missing across all observations but only within sub-samples of the data.

### Missing Completely at Random (MCAR)

In the MCAR situation, the data is missing across all observations regardless of the expected value or other variables.

### Missing Not at Random (MNAR)

The MNAR category applies when the missing data has a structure to it. In other words, there appear to be reasons the data is missing.

#### Deletion

There are two primary methods for deleting data when dealing with missing data: list wise and dropping variables.

### Listwise

In this method, all data for an observation that has one or more missing values are deleted. The analysis is run only on observations that have a complete set of data. If the data set is small, it may be the most efficient method to eliminate those cases from the analysis.

## Pairwise

Pairwise deletion assumes data are missing completely at random (MCAR), but all the cases with data, even those with missing data, are used in the analysis. Pairwise deletion allows data scientists to use more of the data.

### Imputation

When data is missing, it may make sense to delete data, as mentioned above. However, that may not be the most effective option. Instead of deletion, data scientists have multiple solutions to impute the value of missing data.

## Mean, Median and Mode

This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, data scientists can calculate the mean or median of the existing observations.

### Q12What is A/B testing?

Ans:-A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better.

A/B testing, also known as split testing, is a marketing technique that involves comparing two versions of a web page or application to see which performs better. These variations, known as A and B, are presented randomly to users. A portion of them will be directed to the first version, and the rest to the second.

### Below are eight rules to consider when A/B testing.

- Hypothesis. Every test starts with a hypothesis that you're trying to prove or refute. ...
- One Variable. ...
- Clear and Aligned Success Metric. ...
- Volume and Statistical Significance. ...
- Test Group and Splits. ...
- Randomization. ...
- Always be Testing but Apply Common Sense. ...
- Documentation.

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

Ans:- Mean imputation does not preserve the relationships among variables. True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased.

**Mean Imputation Leads to an Underestimate of Standard Errors**

A second reason is applies to any type of single imputation. Any statistic that uses the imputed data will have a standard error that's too low. In other words, we get the same mean from mean-imputed data that we would have gotten without the imputations. And yes, there are circumstances where that mean is unbiased. Even so, the standard error of that mean will be too small. Because the imputations are themselves estimates, there is some error associated with them. But our statistical software doesn't know that. It treats it as real data.

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

**Ans:-** 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 we are using to predict the other variable's value is called the independent variable.

Simple linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line. Both variables should be quantitative.

We can perform the linear regression method in a variety of programs and environments, including:

- R linear regression
- MATLAB linear regression
- Sklearn linear regression
- Linear regression Python
- Excel linear regression

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

**Ans:-** Statistics is the branch of mathematics that deals with data. Data is a collection of values. For most of what we do, it will be numerical data, but it can also take other forms. A collection of data is often referred to as a data set or set of data¸ but other words such as a list or simply collection are also often used.

Two branches, descriptive statistics and inferential statistics, comprise the field of statistics.

**Descriptive statistics**

Descriptive statistics is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically

The basic aim of descriptive statistics is to 'present the data' in an understandable way. If you simply write down every piece of data, it means little to someone who sees it; it needs to be summarised. Imagine if, on the TV news, they listed on the screen the votes of every single person interviewed by a polling company; it would just be a huge list of parties, and you couldn't arrive at any meaningful conclusion.

**Inferential statistics**

Inferential statistics is the aspect that deals with making conclusions about the data. Inferential statistics describe the many ways in which statistics derived from observations on samples from study populations can be used to deduce whether or not those populations are truly different.

Statistics is essentially the study of data. It is used in a huge variety of areas; virtually any subject will need some element of data analysis and study. there are various aspects to statistics: the actual data collection, the presentation of the data (descriptive statistics), and the conclusions that can be drawn (inferential statistics).