# WORKSHEET

## 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 0.
a) True b) False

**Answer:** a) True

2. 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?
a) Central Limit Theorem b) Central Mean Theorem
c) Centroid Limit Theorem d) All of the mentioned

**Answer:** a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
a) Modeling event/time data b) Modeling bounded count data
c) Modeling contingency tables d) All of the mentioned

**Answer:** b) Modeling bounded count data

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

**Answer:**
a) The exponent of a normally distributed random variables follows what is called the log- normal distribution

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

**Answer:** c) Poisson

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

**Answer:** b) False

7. 1. Which of the following testing is concerned with making decisions using data?
a) Probability b) Hypothesis
c) Causal d) None of the mentioned

**Answer:** b) Hypothesis

8. 4. Normalized data are centered at______and have units equal to standard deviations of the original data.
a) 0 b) 5
c) 1 d) 10

**Answer:** a) 0

9. Which of the following statement is incorrect with respect to outliers?
a) Outliers can have varying degrees of influence

b) Outliers can be the result of spurious or real processes
c) Outliers cannot conform to the regression relationship
d) None of the mentioned

**Answer:** c) Outliers cannot conform to the regression relationship

WORKSHEET

**Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.**
10. What do you understand by the term Normal Distribution?

**Answer:**
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.
A graphical representation of a normal distribution is sometimes called a bell curve because of its flared shape. The precise shape can vary according to the distribution of the population but the peak is always in the middle and the curve is always symmetrical. In a normal distribution, the mean, mode and median are all the same.

Normal distribution curves are sometimes designed with a histogram inside the curve. The graphs are commonly used in mathematics, statistics and corporate data analytics.

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

**Answer:** Missing data can be anything from missing sequence, incomplete feature, files missing, information incomplete, data entry error etc. Most datasets in the real world contain missing data. Before you can use data with missing data fields, you need to transform those fields to be used for analysis and modelling. Like many other aspects of data science, this too may actually be more art than science. Understanding the data and the domain from which it comes is very important.

Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data/information of the dataset. These techniques are used because removing the data from the dataset every time is not feasible and can lead to a reduction in the size of the dataset to a large extend, which not only raises concerns for biasing the dataset but also leads to incorrect analysis.

Below are the imputation techniques highly recommended**:**

1. Mean imputation.
2. Substitution.
3. Hot deck imputation.
4. Cold deck imputation.
5. Regression imputation.
6. Stochastic regression imputation.
7. Interpolation and extrapolation.

12. What is A/B testing?

**Answer:** A/B testing, also known as split testing, refers to a randomized experimentation process 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 drive business metrics.

Essentially, A/B testing eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to ‘control’ or the original testing variable. Whereas B refers to ‘variation’ or a new version of the original testing variable.

13. Is mean imputation of missing data acceptable practice?

**Answer:** The process of replacing null values in a data collection with the data’s mean is known as mean imputation.
Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics?

**Answer:**

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.

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. 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, and

(3) Trend forecasting.

15. What are the various branches of statistics?

**Answer:**

There are two main branches of statistics

- Inferential Statistic.

- Descriptive Statistic.

Inferential Statistics:

Inferential statistics used to make inference and describe about the population. These stats are more useful when its not easy or possible to examine each member of the population.

Descriptive Statistics:

Descriptive statistics are used to get a brief summary of data. You can have the summary of data in numerical or graphical form.