STATISTICS WORKSHEET-1
(ANSWERS)

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

   ANS.) 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

   ANS.) A

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

   ANS.) B

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

   ANS.) D

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

   ANS.) C

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

   ANS.) B (It does not change CLT)

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

   ANS.) B

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

   ANS.) A

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

   ANS.) C

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

10. 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.

- It is always normal irrespective of sample size.

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

ANS.) A few of the well known attempts to deal with missing data include:

Hot deck and Cold deck imputation - Hot deck is where a missing value was imputed from a randomly selected similar record. The term "hot deck" dates back to the storage of data on Punched cards, and indicates that the information donors come from the same dataset as the recipients. The stack of cards was "hot" because it was currently being processed.

Cold-deck imputation, by contrast, selects donors from another dataset. Due to advances in computer power, more sophisticated methods of imputation have generally superseded the original random and sorted hot deck imputation techniques. It is a method of replacing with response values of similar items in past surveys. It is available in surveys that measure time intervals.

Listwise and pairwise deletion – By far, the most common means of dealing with missing data is listwise deletion (also known as complete case), which is when all cases with a missing value are deleted.

Pairwise deletion (or "available case analysis") involves deleting a case when it is missing a variable required for a particular analysis, but including that case in analyses for which all required variables are present. When pairwise deletion is used, the total N for analysis will not be consistent across parameter estimations. Because of the incomplete N values at some points in time, while still maintaining complete case comparison for other parameters, pairwise deletion can introduce impossible mathematical situations such as correlations that are over 100%.

Mean imputation - Another imputation technique involves replacing any missing value with the mean of that variable for all other cases

Non-negative matrix factorization - (NMF) can take missing data while minimizing its cost function, rather than treating these missing data as zeros that could introduce biases

Regression imputation – Regression imputation has the opposite problem of mean imputation.

Last observation carried forward - It involves sorting a dataset according to any of a number of variables, thus creating an ordered dataset.

Stochastic imputation - A linear regression model is estimated on the basis of observed values in the target variable Y and some explanatory variables X. The model is used to predict values for the missing cases in Y. Missing values of Y are then replaced on the basis of these predictions.

Multiple imputation - Just as there are multiple methods of single imputation, there are multiple methods of multiple imputation as well. One advantage that multiple imputation has over the single imputation and complete case methods is that multiple imputation is flexible and can be used in a wide variety of scenarios. Multiple imputation can be used in cases where the data are missing completely at random, missing at random, and even when the data are missing not at random.

12. What is A/B testing?

ANS.) The A/B testing process can be simplified as follows:

1. You start the A/B testing process by making a claim (hypothesis).
2. You launch your test to gather statistical evidence to accept or reject a claim (hypothesis) about your website visitors.
3. The final data shows you whether your hypothesis was correct, incorrect or inconclusive.

13. Is mean imputation of missing data acceptable practice?

ANS.) NO, because

- Mean imputation does not preserve the relationship among variables.
- Mean imputation leads to an underestimate of standard errors.

14. What is linear regression in statistics?

ANS.) 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?

ANS.) The two main branches of statistics are descriptive statistics and inferential statistics.

Descriptive Statistics - Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data. Descriptive statistics do not, however, allow us to make conclusions beyond the data we have analysed or reach conclusions regarding any hypotheses we might have made. They are simply a way to describe our data.

Descriptive statistics are very important because if we simply presented our raw data it would be hard to visualize what the data was showing, especially if there was a lot of it. Descriptive statistics therefore enables us to present the data in a more meaningful way, which allows simpler interpretation of the data. For example, if we had the results of 100 pieces of students' coursework, we may be interested in the overall performance of those students.

We would also be interested in the distribution or spread of the marks. Descriptive statistics allow us to do this. How to properly describe data through statistics and graphs is an important topic and discussed in other Laerd Statistics guides. Typically, there are two general types of statistic that are used to describe data:

Measure of Central Tendency: these are ways of describing the central position of a frequency distribution for a group of data. In this case, the frequency distribution is simply the distribution and pattern of marks scored by the 100 students from the lowest to the highest. We can describe this central position using a number of statistics, including the mode, median, and mean. You can learn more in our guide: Measures of Central Tendency.

Measures of spread: these are ways of summarizing a group of data by describing how spread out the scores are. For example, the mean score of our 100 students may be 65 out of 100. However, not all students will have scored 65 marks. Rather, their scores will be spread out. Some will be lower and others higher. Measures of spread help us to summarize how spread out these scores are. To describe this spread, a number of statistics are available to us, including the range, quartiles, absolute deviation, variance and standard deviation.

When we use descriptive statistics it is useful to summarize our group of data using a combination of tabulated description (i.e., tables), graphical description (i.e., graphs and charts) and statistical commentary (i.e., a discussion of the results).

Inferential Statistics - Inferential statistics arise out of the fact that sampling naturally incurs sampling error and thus a sample is not expected to perfectly represent the population. The methods of inferential statistics are (1) the estimation of parameter(s) and (2) testing of statistical hypothesis