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.
   1. True
   2. False
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?
   1. Central Limit Theorem
   2. Central Mean Theorem
   3. Centroid Limit Theorem
   4. All of the mentioned
3. Which of the following is incorrect with respect to use of Poisson distribution?
   1. Modeling event/time data
   2. Modeling bounded count data
   3. Modeling contingency tables
   4. All of the mentioned
4. Point out the correct statement.
   1. The exponent of a normally distributed random variables follows what is called the log- normal distribution
   2. Sums of normally distributed random variables are again normally distributed even if the variables are dependent
   3. The square of a standard normal random variable follows what is called chi-squared distribution
   4. All of the mentioned
5. random variables are used to model rates.
   1. Empirical
   2. Binomial
   3. Poisson
   4. All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   1. True
   2. False
7. 1. Which of the following testing is concerned with making decisions using data?
   1. Probability
   2. Hypothesis
   3. Causal
   4. None of the mentioned
8. 4. Normalized data are centered at and have units equal to standard deviations of the original data.
   1. 0
   2. 5
   3. 1
   4. 10
9. Which of the following statement is incorrect with respect to outliers?
   1. Outliers can have varying degrees of influence
   2. Outliers can be the result of spurious or real processes
   3. Outliers cannot conform to the regression relationship
   4. None of the mentioned

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

1. What do you understand by the term Normal Distribution?

Ans. The normal distribution is the most important probability distribution in [statistics](https://statisticsbyjim.com/glossary/statistics/) because it fits many natural phenomena. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution. It is also known as the Gaussian distribution and the bell curve.

The normal distribution is a probability function that describes how the values of a variable are distributed. It is a symmetric distribution where most of the observations cluster around the central peak and the probabilities for values further away from the [mean](https://statisticsbyjim.com/glossary/mean/) taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely.

In this blog post, you’ll learn how to use the normal distribution, about its [parameters](https://statisticsbyjim.com/glossary/parameter/), and how to calculate Z-scores to standardize your data and find probabilities.

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

Ans. Missing information is an unavoidable aspect of data analysis. For example, responses may be missing to items on survey instruments intended to measure cognitive and affective factors. Various imputation methods have been developed and used for treatment of datasets containing missing data. Some popular methods are listed below.

*(1) Listwise Deletion*. Listwise deletion (LD) involves the removal of all individuals with incomplete responses for any items. However, LD reduces the effective sample size (sometimes greatly, resulting in large amounts of missing data), which can, in turn, reduce statistical power for hypothesis testing to unacceptably low levels. LD assumes that the data are MCAR (i.e., their omission is unrelated to all measured variables). When the MCAR assumption is violated, as is often the case in real research settings, the resulting estimates will be biased.

*(2) Zero Imputation*. When data are omitted as incorrect, the zero imputation method is used, in which missing responses are assigned an incorrect value (or zero in the case of dichotomously scored items).

*(3) Mean Imputation*. In this method, the mean of all values within the same attribute is calculated and then imputed in the missing data cells. The method works only if the attribute examined is not nominal.

*(4) Multiple Imputations*. Multiple imputations can incorporate information from all variables in a dataset to derive imputed values for those that are missing. This method has been shown to be an effective tool in a variety of scenarios involving missing data [11], including incomplete item responses [12].

*(5) Regression Imputation*. The linear regression function is calculated from the values within the same attribute and then used as the dependent variable. The other attributes (except the decision attribute) are then used as independent variables. Then the estimated dependent variable is imputed in the missing data cells. This method works only if all considered attributes are not nominal.

*(6) Stochastic Regression Imputation*. Stochastic regression imputation involves a two-step process in which the distribution of relative frequencies for each response category for each member of the sample is first obtained from the observed data.

In this paper, the details of the seven imputation methods used herein are as follows.

*(i) Listwise Deletion*. All instances are deleted that contain more than one missing cell in their attributes.

*(ii) Mean Imputation*. The missing values from each attribute (column or feature) are replaced with the mean of all known values of that attribute. That is, let  be the th missing attribute of the th instance, which is imputed bywhere  is a set of indices that are not missing in  and  is the total number of instances where the th attribute is not missing.

*(iii) Group Mean Imputation*. The process for this method is the same as that for mean imputation. However, the missing values are replaced with the group (or class) mean of all known values of that attribute. Each group represents a target class from among the instances (recorded) that have missing values. Let  be the th missing attribute of the th instance of the th class, which is imputed bywhere  is a set of indices that are not missing in  and  is the total number of instances where the th attribute of the th class is not missing.

*(iv) Predictive Mean Imputation*. In this method, the functional relationship between multiple input variables and single or multiple target variables of the given data is represented in the form of a linear equation. This method sets attributes that have missing values as dependent variables and other attributes as independent variables in order to allow prediction of missing values by creating a regression model using those variables. For a regression target , the MLR equation with  predictors and  training instances can be written as

This can be rewritten in matrix form such that , and the coefficient  can be obtained explicitly by taking a derivative of the squared error function as follows:

*(v) Hot-Deck*. This method is the same in principle as case-based reasoning. In order for attributes that contain missing values to be utilized, values must be found from among the most similar instances of nonmissing values and used to replace the missing values. Therefore, each missing value is replaced with the value of an attribute with the most similar instance as follows:where  is the standard deviation of the th attribute which is not missing.

*(vi) -NN*. Attributes are found via a search among nonmissing attributes using the 3-NN method. Missing values are imputed based on the values of the attributes of the  most similar instances as follows:where  is the index set of the th nearest neighbors of  based on the nonmissing attributes and  is a kernel function that is proportional to the similarity between the two instances  and  ().

*(vii) -Means Clustering*. Attributes are found through formation of -clusters from nonmissing data, after which missing values are imputed. The entire dataset is partitioned into  clusters by maximizing the homogeneity within each cluster and the heterogeneity between clusters as follows:where  is the centroid of  and  is the union of all clusters (). For a missing value , the mean value of the attribute for the instances in the same cluster with  is imputed thus as follows:

**The following are common methods:**

* Mean imputation. Simply calculate the mean of the observed values for that variable for all individuals who are non-missing. ...
* Substitution. ...
* Hot deck imputation. ...
* Cold deck imputation. ...
* **Regression** imputation. ...
* Stochastic **regression** imputation. ...
* Interpolation and extrapolation.

1. What is A/B testing?

Ans. [A/B testing](https://vwo.com/testing/ab-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.

A/B testing is a method of gathering insight to aid in optimization. It involves testing an original design (A) against an alternate version of that design (B) to see which performs better. That original design is also known as “the control” and the alternate version is known as a “variation.”

1. Is mean imputation of missing data acceptable practice?

Ans. Missing data is a common problem in practical data analysis. They are simply observations that we intend to make but did not. In datasets, missing values could be represented as ‘?’, ‘nan’, ’N/A’, blank cell, or sometimes ‘-999’, ’inf’, ‘-inf’. The aim of this tutorial is to provide an introduction of missing data and describe some basic methods on how to handle them.

It is a non-standard, but a fairly flexible **imputation** algorithm. It uses RandomForest at its core to predict the **missing data**. It can be applied to both continuous and categorical variables which makes it advantageous over other **imputation** algorithms.

1. What is linear regression in statistics?

Ans. ***Linear regression***quantifies the relationship between one or more *predictor variable(s)*and one *outcome variable.*Linear regression is commonly used for predictive analysis and modeling. For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable).  Linear regression is also known as *multiple regression*,*multivariate regression*,*ordinary least squares (OLS)*, and *regression*. This post will show you examples of linear regression, including an example of *simple linear regression* and an example of *multiple linear regression*.

1. What are the various branches of statistics?

Ans. Statistics is the main branch of [mathematics](https://statanalytica.com/math-assignment-help). Used to perform different operations, i.e., Data collection, organization, analysis, and so on. In other words, statistics are a form of mathematical analysis that uses quantitative models to give a set of experimental data or studies of real life. Statistics examine the methodology for collecting, reviewing, analyzing, and making data conclusions.

**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](https://statanalytica.com/data-analysis-assignment-help) on different studies.

**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](https://statanalytica.com/probability-assignment-help) 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

