Statistics-WORKSHEET-1

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

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?

b) Central Mean Theorem

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

c) Modeling contingency tables

4. Point out the correct statement

d) All of the mentioned

5. --- random variables are used to model rates.

c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.

b) False

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

b) Hypothesis

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

c) 1

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

d) None of the mentioned

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

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. As with any probability distribution, the parameters for the normal distribution define its shape and probabilities entirely.

The normal distribution has two parameters, the mean and standard deviation. The normal distribution does not have just one form. Instead, the shape changes based on the [parameter](https://statisticsbyjim.com/glossary/parameter/) values.

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

Missing data can be handled by:

* Deleting Rows with missing values.
* Impute missing values for continuous variable.
* Impute missing values for categorical variable.
* Other Imputation Methods.
* Using Algorithms that support missing values.
* Prediction of missing values.

Common imputation statistics include:

1. The column mean value.
2. The column median value.
3. The column mode value.
4. A constant value.

12. What is A/B testing?

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment. We can use random experiments, or can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

13. Is mean imputation of missing data acceptable practice?

When the dataset we want to use for Machine Learning contains missing data, we normally substitute a mean for numerical feature and mode for categorical ones. Sometimes, just insert 0's or discard the data and proceed to the training of the model. But by using a mean or mode can significantly reduce the model’s accuracy and bias the results.

The following reasons are:

* + Mean and Mode ignore feature correlations
  + Mean reduces a variance of the data

14. What Is Linear Regression in Statistics?

In [statistics](https://en.wikipedia.org/wiki/Statistics), Linear Regression is a [linear](https://en.wikipedia.org/wiki/Linearity) approach to modelling the relationship between a [scalar](https://en.wikipedia.org/wiki/Scalar_(mathematics)) response and one or more explanatory variables known as [dependent and independent variables](https://en.wikipedia.org/wiki/Dependent_and_independent_variables). The case of one explanatory variable is called [simple linear regression](https://en.wikipedia.org/wiki/Simple_linear_regression); for more than one, the process is called multiple linear regressions. In linear regression, the relationships are modelled using [linear predictor functions](https://en.wikipedia.org/wiki/Linear_predictor_function) whose unknown model [parameters](https://en.wikipedia.org/wiki/Parameters) are [estimated](https://en.wikipedia.org/wiki/Estimation_theory) from the [data](https://en.wikipedia.org/wiki/Data). Such models are called [linear models](https://en.wikipedia.org/wiki/Linear_model). Like all forms of [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis), linear regression focuses on the [conditional probability distribution](https://en.wikipedia.org/wiki/Conditional_probability_distribution) of the response given the values of the predictors, rather than on the [joint probability distribution](https://en.wikipedia.org/wiki/Joint_probability_distribution) of all of these variables, which is the domain of [multivariate analysis](https://en.wikipedia.org/wiki/Multivariate_analysis).

15. What are the various branches of statistics?

The two main branches of statistics are Descriptive statistics and Inferential statistics. Both of these are employed in scientific analysis of data and both are equally important.

1. Descriptive statistics: It implies a simple quantitative summary of a data set that has been collected. It helps us understand the experiment or data set in detail and tells us everything we need to put the data in perspective.
2. Inferential statistics: It is the attempt to apply the conclusions that have been obtained from one experimental study to more general populations. This means inferential statistics tries to answer questions about populations and samples that have not been tested in the given experiment.