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.

Ans=a) True

1. 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?

Ans=a) Central Limit Theorem

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

Ans=b) Modeling bounded count data

1. Point out the correct statement.

Ans=c) The square of a standard normal random variable follows what is called chi-squared distribution

1. \_\_\_random variables are used to model rates.

Ans=c) Poisson

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

Ans=b) False

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

Ans =b) Hypothesis

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

Ans =a) 0

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

Ans =c) Outliers cannot conform to the regression relationship

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 =a) The normal distribution is also known as Gaussian distribution , is a probability distribution

I .e. symmetric about the mean, showing 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.

b) The normal distribution is a probability distribution that describes many common datasets in the real world. It is the most common type of distribution, and it arises naturally in statistics through random sampling techniques.

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

Ans=Imputation vs Removing Data

**1. Missing at Random (MAR):**Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data

**2. Missing Completely at Random (MCAR):** The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.

**3. Missing not at Random (MNAR):**Two possible reasons are that the missing value depends on the hypothetical value.

Deletion

**Listwise=**Listwise deletion (complete-case analysis) removes all data for an observation that has one or more missing values. Particularly if the missing data is limited to a small number of observations, you may just opt to eliminate those cases from the analysis. However in most cases, it is often disadvantageous to use listwise deletion. This is because the assumptions of MCAR (Missing Completely at Random) are typically rare to support. As a result, listwise deletion methods produce biased parameters and estimates.

**Pairwise=**pairwise deletion analyses all cases in which the variables of interest are present and thus maximizes all data available by an analysis basis. A strength to this technique is that it increases power in your analysis but it has many disadvantages. It assumes that the missing data are MCAR. If you delete pairwise then you’ll end up with different numbers of observations contributing to different parts of your model, which can make interpretation difficult.

**Dropping Variables**In my opinion, it is always better to keep data than to discard it. Sometimes you can drop variables if the data is missing for more than 60% observations but only if that variable is insignificant. Having said that, imputation is always a preferred choice over dropping variables

12.What is A/B testing?

Ans=A/B testing is the process of comparing two variations of a page element, usually by testing users' response to variant A vs variant B, and concluding which of the two variants is more effective.

A/B testing is also known as split testing. It is a process of showing two variants of the same web page to different segments of websites visitors at the same time and comparing which variant drives more conversions.

13. Is mean imputation of missing data acceptable practice?

Ans= i) It is a non-standard, but a fairly flexible **imputation** algorithm. It uses Random Forest 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.

ii) Bad practice in general

If just estimating means: mean imputation preserves the mean of the observed data

Leads to an underestimate of the standard deviation

Distorts relationships between variables by “pulling” estimates of the correlation toward zero.

14. What is linear regression in statistics?

Ans=) Linear regression attempts to model the relationship between the two variable by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

ii) Linear regression is a basic and commonly used type of predictive analysis. These regression estimates are used to explain the relationship between one dependent variable and one or more independent variable.

15. What are the various branches of statistics?

Ans=The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed by in scientific analysis of data and both are equally important for the student of statistics.

Descriptive Statistics

Inferential statistics

1)Descriptive statistics=Descriptive [statistics](https://www.investopedia.com/terms/s/statistics.asp) are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median and mode, while measures of [variability](https://www.investopedia.com/terms/v/variability.asp) include [standard deviation](https://www.investopedia.com/ask/answers/042415/what-difference-between-standard-error-means-and-standard-deviation.asp), variance, minimum and maximum variables, and [kurtosis](https://www.investopedia.com/terms/k/kurtosis.asp) and skewness.

2)Inferential Statistics=Inferential statistics use statistical models to help you compare your sample data to other samples or to previous research. Most research uses statistical models called the[Generalized Linear model](https://www.statisticshowto.com/general-linear-model-glm/)and include [Student’s t-tests](https://www.statisticshowto.com/probability-and-statistics/t-test/), [ANOVA (Analysis of Variance](https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/anova/)), [regression](https://www.statisticshowto.com/probability-and-statistics/regression-analysis/)analysis and various other models that result in straight-line (“[linear](https://calculushowto.com/linear-function/#definition)“) probabilities and results.