**Statistics– WORKSHEET**

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

**a) True**

b) False

ANS- A

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?

**a) Central Limit Theorem**

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

ANS- A

1. 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

1. 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**

1. random variables are used to model rates.

a) Empirical

b) Binomial

**c) Poisson**

d) All of the mentioned

ANS- C

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

a) True

**b) False**

**ANS- B**

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

1. . 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

1. 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

1. 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

The normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses. The standard normal distribution has two parameters: the mean and the standard deviation. For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations.

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

**Ans-** The best possible method of handling the missing data is to prevent the problem by well-planning the study and collecting the data carefully . The following are suggested to minimize the amount of missing data in the clinical research .

First, the study design should limit the collection of data to those who are participating in the study. This can be achieved by minimizing the number of follow-up visits, collecting only the essential information at each visit, and developing the user friendly case-report forms.

Second, before the beginning of the clinical research, a detailed documentation of the study should be developed in the form of the manual of operations, which includes the methods to screen the participants, protocol to train the investigators and participants, methods to communicate between the investigators or between the investigators and participants, implementation of the treatment, and procedure to collect, enter, and edit data.

Third, before the start of the participant enrollment, a training should be conducted to instruct all personnel related to the study on all aspects of the study, such as the participant enrollment, collection and entry of data, and implementation of the treatment or intervention .

Fourth, if a small pilot study is performed before the start of the main trial, it may help to identify the unexpected problems which are likely to occur during the study, thus reducing the amount of missing data.

Fifth, the study management team should set a priori targets for the unacceptable level of missing data. With these targets in mind, the data collection at each site should be monitored and reported in as close to real-time as possible during the course of the study.

Sixth, study investigators should identify and aggressively, though not coercively, engage the participants who are at the greatest risk of being lost during follow-up.

Finally, if a patient decides to withdraw from the follow-up, the reasons for the withdrawal should be recorded for the subsequent analysis in the interpretation of the results.

The simplest imputation method is replacing missing values with the mean or median values of the dataset at large, or some similar summary statistic. This has the advantage of being the simplest possible approach, and one that doesn't introduce any undue bias into the dataset.

1. What is A/B testing?

ANS- A/B testing is the act of running a simultaneous experiment between two or more variants of a page **to see which one performs the best**.

Imagine, for instance, that you want to test your hypothesis that one headline will generate more leads than another. Sure, you could just make the change and cross your fingers. But what if you’re wrong? Mistakes can get costly.

By sending half your traffic to one version of the page and half to another, you can first **gather evidence about which one works best** before you commit to the change.

Essentially, A/B testing lets you play scientist—and **make decisions based on data** about how people actually behave when they hit your page.

1. Is mean imputation of missing data acceptable practice?

ANS - 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- In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

1. 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 in scientific analysis of data and both are equally important for the student of statistics.