STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0
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?
Ans-> Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
Ans -> Modeling bounded count data

4. Point out the correct statement.
Ans -> All of the mentioned

5. _____ random variables are used to model rates
Ans -> Poisson

6. Usually replacing the standard error by its estimated value does change the CLT
Ans -> False

7. Which of the following testing is concerned with making decisions using data?
Ans-> Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
Ans-> 0

9. Which of the following statement is incorrect with respect to outliers?
Ans :- Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?
Ans -> A normal distribution has a probability distribution that is centered around the mean. This means that the distribution has more data around the mean. The data distribution decreases as you move away from the center. The resulting curve is symmetrical about the mean and forms a bell-shaped distribution.

11. How do you handle missing data? What imputation techniques do you recommend?
Ans -> Missing data can be dealt with in a variety of ways. I believe the most common reaction is to

ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programme will make the decision for you.

Your application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea.

Another common strategy among those who pay attention is imputation. Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values.

And how would you choose that estimate? The following are some of the most prevalent methods:

Mean imputation

Calculate the mean of the observed values for that variable for all non-missing people. It has the advantage of maintaining the same mean and sample size, but it also has a slew of d rawbacks. Almost all of the methods described below are superior to mean imputation.

Substitution

Assume the value from a new person who was not included in the sample. To put it another way, pick a new subject and employ their worth instead.

Hot deck imputation

value picked at random from a sample member who has comparable values on other variables. To put it another way, select all the sample participants who are comparable on other factors, then choose one of their missing variable values at random.

One benefit is that you are limited to just feasible values. In other words, if age is only allowed to be between 5 and 10 in your research, you will always obtain a value between 5 and 10.Another factor is the random element, which introduces some variation. For exact standard errors, this is crucial.

Cold deck imputation

A value picked deliberately from an individual with similar values on other variables. In most aspects, this is comparable to Hot Deck, but without the random variance. As an example, under the same experimental condition and block, you can always select the third individual.

Regression imputation
The result of regressing the missing variable on other factors to get a predicted value. As a result,

instead of utilising the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

Stochastic regression imputation

The predicted value of a regression plus a random residual value. This has all of the benefits of regression imputation plus the random component's benefits. The majority of multiple imputation is based on stochastic regression imputation.

Interpolation and extrapolation

An estimate based on other observations made by the same person. It generally only works with data that is collected over time. Proceed with caution, though. For a variable like height in children– one that cannot be reduced through time–interpolation would make more sense. Extrapolation entails estimating beyond the data's true range, which necessitates making more assumptions than is necessary.

12. What is A/B testing?
Ans -> A/B testing is one of the most important concepts in data science and in the tech world in general because it is one of the most effective methods in making conclusions about any hypothesis one may have. It's important that you understand what A/B testing is and how it generally works.

13. Is mean imputation of missing data acceptable practice?
Ans -> Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.
Mean imputation (MI) is one such method in which the mean of the observed values for each variable is computed and the missing values for that variable are imputed by this mean. This method can lead into severely biased estimates even if data are MCAR (Missing completely at random)

14. What is linear regression in statistics?
Ans -> Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

15. What are the various branches of statistics?

Ans -> There are three real branches of statistics: data collection, descriptive statistics and inferential statistics.

Data collection :- Data collection is a process of gathering information from all the relevant sources to find a solution to the research problem. It helps to evaluate the outcome of the problem. The data collection methods allow a person to conclude an answer to the relevant question.

Descriptive statistics :- Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread).

Inferential statistics :- Inferential statistics is mainly used to derive estimates about a large group (or population) and draw conclusions on the data based on hypotheses testing methods. Inferential statistics uses sample data because it is more cost-effective and less tedious than collecting data from an entire population.