PROJECT - INTERNSHIP-26

STATISTICAL WORKSHEET 1

1. Bernoulli random variables take (only) the values 1 and 0
   = Option (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?
   = Option (A) Central Limit Theorem
3. Which of the following is incorrect with respect to use of Poisson Distribution?
   = Option (A) Modeling event/time data
4. Point out the correct statement:
   = Option (B) Sums of normally distributed random variables are again normally distributed even if the variables are dependent.
5. ……. random variables are used to model rates.
   = Option (C) Poisson
6. Usually replacing the standard error by its estimated value does change the CLT.
   = Option (B) False
7. Which of the following testing is concerned with making decision using data?
   = Option (B) Hypothesis
8. Normalized data are centered at ……. and have units equal to standard deviations of the original data.
   = Option (A) 0
9. Which of the following statement is incorrect with respect to outliers?
   = Option (C) Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?
    A normal distribution of data is one in which the majority of data points are relatively similar, meaning they occur within a small range of values with fewer outliers on the high and low ends of the data range. In a normal distribution, the mean, mean and mode are equal. The total area under the curve should be equal to 1. The normally distributed curve should be symmetric at the centre. There should be exactly half of the values are to the right of the centre and exactly half of the values are to

the left of the centre. In normal distribution data near the mean are more frequent in occurrence than data far from the mean.
All kinds of variables in natural and social sciences are normally or approximately normally distributed. Height, birth weight, reading ability, job satisfaction, or SAT scores are just a few examples of such variables.

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

    Missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. Missing data can occur because of nonresponse: no information is provided for one or more items or for a whole unit ("subject"). Some items are more likely to generate a nonresponse than others: for example items about private subjects such as income. Missingness occurs when participants drop out before the test ends and one or more measurements are missing. Understanding the reasons why data are missing is important for handling the remaining data correctly.

    When data is missing, it may make sense to delete data, as mentioned above. However, that may not be the most effective option. For example, if too much information is discarded, it may not be possible to complete a reliable analysis. Or there may be insufficient data to generate a reliable prediction for observations that have missing data.

    Instead of deletion, data scientists have multiple solutions to impute the value of missing data. Depending why the data are missing, imputation methods can deliver reasonably reliable results.

    Mean, Median and Mode

    This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, data scientists can calculate the mean or median of the existing observations. However, when there are many missing variables, mean or median results can result in a loss of variation in the data. This method does not use time-series characteristics or depend on the relationship between the variables.

    Time-Series Specific Methods

Another option is to use time-series specific methods when appropriate to impute data.The time series methods of imputation assume the adjacent observations will be like the missing data. These methods work well when that assumption is valid. However, these methods won't always produce reasonable results, particularly in the case of strong seasonality.

MULTIPLE IMPUTATION

Multiple imputation is considered a good approach for data sets with a large amount of missing data. Instead of substituting a single value for each missing data point, the missing values are exchanged for values that encompass the natural variability and uncertainty of the right values. Using the imputed data, the process is repeated to make multiple imputed data sets. Each set is then analyzed using the standard analytical procedures, and the multiple analysis results are combined to produce an overall result.

The various imputations incorporate natural variability into the missing values, which creates a valid statistical inference. Multiple imputations can produce statistically valid results even when there is a small sample size or a large amount of missing data.

12. What is A/B testing?

A/B 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 eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to 'control' or the original testing variable. Whereas B refers to 'variation' or a new version of the original testing variable.

The version that moves your business metric(s) in the positive direction is known as the 'winner.' Implementing the changes of this winning variation on your tested page(s) / element(s) can help optimize your website and increase business

13. Is mean imputation of missing data acceptable practise?

Mean imputation (also called mean substitution) really ought to be a last way out to the problem. It's a popular solution to missing data, despite its drawbacks. Mainly because it's easy. It can be really painful to lose a large part of the sample you so carefully collected, only to have little power.

14. What is Linear Regression in statistics?

In statistics, linear regression is a linear approach for modelling the relationship between dependent and independent variables. The case of one explanatory variable is called *simple linear regression*; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

15. What are various branches of Statistics?

- Descriptive Statistics. Descriptive statistics is the first part of statistics that deals with the collection of data.
- 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.