Statistics Worksheet-1

A.1) Option a- True

A.2) Option a- Central Limit Theorem

A.3) Option b- Modeling bounded count data

A.4) Option d- All of the mentioned

A.5) Option c- Poisson

A.6) Option b- False

A.7) Option b- Hypothesis

A.8) Option a- 0

A.9) Option c- Outliers cannot conform to the regression relationship

A.10) Normal Distribution- A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the mean of the distribution.
The normal distribution is also known as a Gaussian distribution or probability bell curve. It is symmetric about the mean and indicates that values near the mean occur more frequently than the values that are farther away from the mean.
Graphically, a normal distribution is a bell curve because of its flared shape. The precise shape can vary according to the distribution of the values within the population. The population is the entire set of data points that are part of the distribution.
Regardless of its exact shape, a normal distribution bell curve is always symmetrical about the mean. A symmetrical distribution means that a vertical dividing line drawn through the maximum/mean value will produce two mirror images on either side of the line, in which half the population is less than the mean and half is greater. However, the reverse is not always true; that is, not all symmetrical distributions are normal. In the bell curve, the peak is always in the middle, and the mean, mode and median are all the same.

A.11) Missing data can be handle 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 our statistical programed will make the decision for us.
Our 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 analyzing the entire data set as if the imputed values were the true observed values.
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 drawbacks. 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

A 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 we are limited to just feasible values. In other words, if age is only allowed to be between 5 and 10 in our research, we 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, we 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 utilizing the mean, we'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.

Single or Multiple Imputation

Single and multiple imputation are the two forms of imputation. When people say imputation, they usually mean single.

The term "single" refers to the fact that we only use one of the seven methods to estimate the missing number outlined above.

It's popular since it's simple to understand and generates a sample with the same number of observations as the complete data set.

When listwise deletion eliminates a considerable amount of the data set, single imputation appears to be a tempting option. It does, however, have certain restrictions.

Unless the data is Missing Completely at Random, certain imputation processes, such as means, correlations, and regression coefficients, result in skewed parameter estimations.

I recommend Single or Multiple Imputation techniques.

A.12) A/B testing is a popular way to test our products and is gaining steam in the data science field.

A/B testing is a type of experiment in which we split our web traffic or user base into two groups, and show two different versions of a web page, app, email, and so on, with the goal of comparing the results.

A.13) Mean imputation of missing data is not a good practice because of mean imputation does not preserve the relationship among variables and mean imputation leads to an underestimate of standard errors.

A.14) Linear regression analysis is used to predict the value of a variable based on the value of another variable.

A.15) Descriptive Statistics and Inferential Statistics

Descriptive Statistics: It is used to describe the features of data and shown or summarize data in the form of table, charts, and graphs. Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis.

Inferential Statistics: It is used to study the relationship among variables in data and helps in making predictions, conclusions, or generalization about the whole population. Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics.