**Internship 26**

**Statistics Worksheet 1**

Question 1. (A) True

Question 2. (A) Central Limit Theorem

Question 3. (B) Modelling Bounded Count Data

Question 4. (D) All of the mentioned.

Question 5. (C) Poisson

Question 6. (B) False

Question 7. (B) Hypothesis

Question 8. (A) 0

Question 9. (C) Outliers cannot conform to the regression relationship.

Question 10. **What is Normal Distribution?**

A normal distribution is the continuous probability distribution with a probability density function that gives you us a symmetrical bell curve.

It is a plot of the probability function of a variable that has maximum data concentrated around one point and a few points taper off symmetrically towards two opposite ends.

A normal distribution has a probability distribution that is centred around the mean. This means that the distribution has more data around the mean. The data distribution decreases as you move away from the centre. The resulting curve is symmetrical about the mean and forms a bell-shaped distribution.

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

Missing data are problematic because, depending on the type, they can sometimes bias our results. This means our results may not be generalizable outside of our study because our data come from an unrepresentative sample. **Missing data**, or missing values, occur when we don’t have data stored for certain variables or participants. Data can go missing due to incomplete data entry, equipment malfunctions, lost files, and many other reasons. In any dataset, there are usually some missing data. In quantitative research, missing values appear as blank cells in your spreadsheet. The reason for the missing data is important to consider, because it helps you determine the type of missing data and what you need to do about it. There are three main types of missing data namely Missing completely at random (MCAR), Missing at random (MAR), Missing not at random (MNAR). MNAR are also called Non-Ignorable while MCAR and MAR are ignorable.

To clean our data, our options usually include accepting, removing, or recreating the missing data. We should consider how to deal with each case of missing data based on our assessment of why the data are missing. Are these data missing for random or non-random reasons? Are the data missing because they represent zero or null values? Was the question or measure poorly designed? Our data can be accepted or left as it is, if it’s MCAR or MAR. However, MNAR data need more complex treatment.

Imputationmeans replacing a missing value with another value based on a reasonable estimate and analysing dataset as if the imputed values are true observed values. We use other data to recreate the missing value for a more complete dataset. We can choose from several imputation methods. Following are the examples of Imputation methods which can be used :

1. **Mean Imputation** :- We calculate the mean of the observed values for that variable for all non-missing values.
2. Substitution method :- Assume the value from a new person who was not included in the sample.
3. **Hot Deck Imputation** :- A value picked at random from a sample member who has comparable values on other variables. In other ways, select all the sample participants who are comparable on other factors, then choose one of their missing variable values at random.
4. **Cold Deck Impu**tation :- A value picked deliberately from an individual with similar values on other variables. In most aspects, this is comparable to Hot Deck Imputation, but without the random variance.
5. **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, We are 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.
6. **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.
7. **Interpolation** :- An estimate based on other observations is made by the same person. It generally only works with data that is collected over time.
8. **Extrapolation** :- This is estimating beyond the data's true range, which necessitates making more assumptions than is necessary. Standard errors are underestimated by all single imputation approaches. Because the imputed observations are estimates, their values have a random error associated with them. However, our programme is unaware of this when we enter that estimate as a data point. As a result, it ignores the additional source of error, resulting in too-small standard errors and p-values. While imputation is straightforward in theory, it is difficult to master in reality. As a result, it isn't perfect, although it may suffice in some circumstances. As a result of multiple imputation, numerous estimates are generated. In multiple imputation, two of the approaches indicated above–hot deck and stochastic regression–work as the better imputation methods. The multiple imputation estimates varies significantly because these two approaches contain a random component. This reintroduces some variance that our program can account for in order to provide reliable standard error estimates for our model.

Question 12**. What is A/B Testing?**

**It** is one of the most popular controlled experiments used to optimize web marketing strategies. It allows decision makers to choose the best design for a website by looking at the analytics results obtained with two possible alternatives A and B. For example, there are two alternative designs A and B. Visitors of a website are randomly served with A and B. Then, data about their activity is collected by web analytics. Given this data, one can apply statistical tests to determine which one of the two designs has better efficiency. Different kinds of metrics like **binomial metrics**, non-binomial metrics, sample size, and sampling distributions require different kinds of statistical tests for computing the efficiency of A/B tests such as null hypothesis, Fisher’s exact test, Pearson’s chi-squared test, z test etc.

Question 13. **Is mean imputation of missing data acceptable practice?**

Mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable. It comes under the category of single imputation.

1. Missing values in your data do not reduce our sample size, as it would be the case with listwise deletion. Since mean imputation replaces all missing values, we can keep our whole database.
2. Mean imputation is very simple to understand and to apply. We can explain the imputation method easily to our audience and everybody with basic knowledge in statistics will get what we’ve done.
3. If the response mechanism is MCAR, the sample mean of our variable is not biased. Mean substitution might be a valid approach, in case that the univariate average of our variables is the only thing we are interested in.

**Following are some drawbacks of Mean Imputation** :

1. Mean substitution leads to bias in multivariate estimates such as correlation or regression coefficients. Values that are imputed by a variable’s mean have, in general, a correlation of zero with other variables. Relationships between variables are therefore biased toward zero.
2. Standard errors and variance of imputed variables are biased. For instance, let’s assume that we would like to calculate the standard error of a mean estimation of an imputed variable. Since all imputed values are exactly the mean of our variable, we would be too sure about the correctness of our mean estimate. In other words, the confidence interval around the point estimation of our mean would be too narrow.
3. If the response mechanism is MAR OR MNAR, even the sample mean of our variable is biased (compare that with point 2 above). Assume that we want to estimate the mean of an income of population and people with high income are less likely to respond; Our estimate of the mean income would be biased downwards.

Hence, we can conclude that mean imputation hampers the quality of our data analysis.

Question 14. **What is Linear Regression in Statistics?**

**Linear Regression** is a field of study which emphasizes on the statistical relationship between two **continuous variables** known as **Predictor**and **Response**variables. When there are more than one predictor variables then it is called multiple linear regression. Regression is the technique which helps us to determine difficult to measure variables with the help of easy to measure variables. Linear Regression is very powerful technique to predict the value of a response variable when there is a linear relationship between two continuous variables.

Predictor variable is often denoted as x and also known as an independent variable. Response variable is often denoted as y and also known as dependent variable.

**Statistical relationship is not exact but we can see the trend using it. This hypothesis is supported by scatter plot. For example, Height and Weight, as height increases, we expect weight to be increased but not in exact accordance with the height, Speed of Car and Mileage, As Speed increases we expect the Mileage to decrease but not perfectly.**

Question 15. **What are the various branches of Statistics?**

**Descriptive Statistics** : Descriptive statistics is the first part of statistics that deals with the collection of data. The statisticians need to be aware of the design and experiments for collection of correct data. They also need to select the correct focus group and eliminate biases. It is also used for doing various analysis on different studies.

Descriptive Statistics has Two measures namely Central tendency measures (like Mean, Median, Mode) and Variability measures to help understand the analysed data.

The variability measures help statisticians to analyse the distribution that is spreading from a specific data set. Examples of Variability measures include quartiles, ranges, variances, and standard deviation.

**Inferential Statistics** : Inferential statistics includes techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population. Inferential statistics works in terms of probability by using descriptive statistics. A statisticians use various techniques for data analysis, drafting, and making conclusions from limited information, that is obtained by taking sample data and testing the reliability of analysis. For example, social experiments that deal with studying a small sample population that helps determine the community’s behaviour. Inferential statistics use following methods and tools like regression analysis, analysis of variance, analysis of co-variance, t-test, correlation analysis etc.