FLIP ROBO – STATISTICS WORKSHEET 1 SOLUTION | STUDENT NAME: OLUWAJOBA FATOLA

1. **a)** True
2. **a)** Central Limit Theorem.
3. **c)** Modeling contingency tables
4. **c)** The square of a standard normal random variable follows what is called chi-squared distribution
5. **c)** Poisson
6. **b)** False
7. **b)** Hypothesis
8. **a)** 0
9. **c)** Outliers cannot conform to the regression relationship
10. The normal distribution, also referred to as the Gaussian distribution or the bell curve, is a symmetric and bell-shaped probability distribution. Its mean (μ) and standard deviation (σ) are used to define it. The data in a normal distribution clusters around the mean, and the standard deviation determines the spread of the data.

    The characteristics of a normal distribution include:
    - **Symmetry**: The distribution is symmetric around the mean, meaning that the data is equally likely to be above or below the mean.
    - **Bell-shaped curve**: The shape of the distribution resembles a bell, with the highest point at the mean and the data tapering off symmetrically on both sides.
    - **Central tendency**: The mean, median, and mode of a normal distribution are all equal and located at the center of the distribution.
    - **Standard deviation**: The standard deviation determines the spread of the data. A larger standard deviation indicates a wider spread, while a smaller standard deviation indicates a narrower spread.

    The normal distribution is widely used in statistics and probability theory due to its mathematical properties and its applicability to many real-world phenomena. It is particularly important in inferential statistics and hypothesis testing, as many statistical tests assume normality of the data. For example, the heights of people, the IQ scores of people, and the test scores of students are all often modelled by normal distributions.

11. There are two main ways to handle missing data:

    A. **Imputation** is the process of replacing missing values with estimated values. There are many different imputation techniques, including:
        - **Mean imputation** replaces missing values with the mean of the observed values.
        - **Median imputation** replaces missing values with the median of the observed values.
        - **Mode imputation** replaces missing values with the mode of the observed values.
        - **Regression imputation** uses a regression model to predict the missing values. A regression model is built using the observed data and the missing values are then imputed using the predicted values from the model.

- **Multiple imputation** uses multiple imputation models to predict the missing values. Each imputed dataset is analysed separately, and the results are combined to obtain the final estimates.
- **Advanced techniques** such as k-nearest neighbours, expectation-maximization imputation, and Bayesian imputation.

B. **Deletion** is the process of removing cases with missing values. There are two main types of deletion:

- **Listwise deletion** removes cases with any missing values.
- **Pairwise deletion** removes cases with missing values only for the variables being analyzed.

The best way to handle missing data depends on the type of data, the amount of missing data, and the purpose of the analysis. In general, imputation is preferred over deletion because it preserves more information. However, imputation can introduce bias if the imputation model is not accurate.

I recommend using multiple imputation if the amount of missing data is significant. Multiple imputation is more accurate than other imputation techniques and it can reduce the bias introduced by deletion.

Here are some additional tips for handling missing data:

- Identify the type of missing data. Missing data can be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). MCAR means that the probability of a value being missing is independent of any other values in the dataset. MAR means that the probability of a value being missing is dependent on other values in the dataset, but not on the value itself. MNAR means that the probability of a value being missing is dependent on the value itself.
- Explore the data. Look at the distribution of missing values and see if there are any patterns. For example, if all the missing values are in one variable, then you may want to consider deleting that variable.
- Use a variety of imputation techniques. Try different imputation techniques and see which one works best for your data.
- Evaluate the results. After you have imputed the missing data, evaluate the results to make sure that they are reasonable.

12. A/B testing, also known as split testing, is a statistical method for comparing two alternative versions of a webpage, marketing campaign, or other piece to see which one works better. It is widely used in marketing and web development to improve conversion rates, user engagement, and other key performance indicators.

In A/B testing, two versions, known as the control group (A) and the treatment group (B), are created. The control group represents the existing version or standard practice, while the treatment group represents a modified version or a new approach. Randomly selected users

or participants are divided into these two groups, and their responses and behaviours are measured and compared.
The objective of A/B testing is to determine whether the changes made in the treatment group result in statistically significant improvements compared to the control group. By comparing the performance metrics of the two groups, such as click-through rates, conversion rates, or revenue, the impact of the changes can be assessed.

A/B testing allows businesses and organizations to make data-driven decisions by testing different variations and identifying the most effective approaches. It helps optimize user experiences, marketing campaigns, website designs, and other factors that impact performance and success. It also helps organisations to improve the performance of the website or app, increase conversions, sales, or leads, and helps to save money on marketing and advertising.

13. Mean data imputation is an acceptable practice in handling missing data as it is a simple and easy method but not always the best option. One of the main concerns with mean imputation is that it can introduce bias into the data. By replacing missing values with the mean, the imputed values tend to cluster around the mean, leading to an underestimation of the true variability in the data. This can affect subsequent analyses and statistical models by distorting the relationships and reducing the accuracy of predictions.

    Additionally, mean imputation assumes that the missing values are missing completely at random (MCAR) or missing at random (MAR), meaning that the probability of missingness is unrelated to the missing values themselves. If the missing values are not MCAR or MAR, mean imputation may introduce additional bias and distort the results. The mean of the observed values may not be representative of the true mean of the population, especially if the missing values are not randomly distributed. If the missing values are correlated with other variables, then imputing the mean can make the relationships between those variables appear stronger or weaker than they actually are.

    In summary, while **mean imputation is a simple and convenient method for handling missing data**, it has limitations and can introduce bias. It is important to carefully consider the nature of the missing data and explore alternative imputation techniques that are more suitable for the specific dataset and research objectives.

14. Linear regression is a statistical method that is used to model the relationship between a dependent variable (target variable) and one or more independent variables (response variable). The dependent variable is the variable that you are trying to predict, and the independent variables are the variables that you are using to predict the dependent variable. The line is represented by a linear equation of the form:

    $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n * X_n$$

    where:

    - Y is the dependent variable

- β_0 is the intercept
- $\beta_1, \beta_2, ..., \beta_n$ are the coefficients
- $X_1, X_2, ..., X_n$ are the independent variables

the goal is to find the best-fitting line that minimizes the differences between the observed data points and the predicted values based on the linear relationship.

In linear regression, the relationship between the dependent variable and the independent variables is assumed to be linear. This means that the dependent variable can be represented as a straight-line function of the independent variables.
Another important assumption is that of independence errors, homoscedasticity (constant variance of errors), and the normality of errors. The validity and the accuracy of the regression result can be affected by the violations of these assumptions.

15. Branches of statistics is represented in the below diagram: