WORKSHEET 1 - STATISTICS

Submitted by - Jwala R, DS0082 (18/01/2023)

1. a) True
2. a) Central Limit Theorem
3. b) Modeling bounded count data
4. d) All of the mentioned
5. c) Poisson
6. b) False
7. b) Hypothesis
8. a) 0
9. c) Outliers cannot conform to the regression relationship

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.

    Let's assume that X is the random variable and that f(x) is the probability density function. In order to determine the probability of the random variable X, it specifies a function that is integrated across the range or interval (x to x + dx) while taking into account values between x and x+dx.

    $f(x) \geq 0 \ \forall \ x \in (-\infty,+\infty)$
    And $-\infty\int+\infty\ f(x) = 1$

    Where, x is the variable, μ is the mean and σ is the standard deviation.

    Eg: Marks scored on the test and Heights of different persons

11. There are several approaches to handle missing data.Most of the time, the programme will delete items in a listwise order. Listwise deletion may or may not be a wise choice, depending on why and how much data was lost.

    Imputation is another way used by people who pay attention. Imputation involves replacing missing values with an estimate and analysing the complete set of data as if the imputed values were the actual observed values.

    Types:

    - Mean imputation - Determine the mean of all non-missing individuals' observed values for that variable. It has the benefit of keeping the mean and sample size constant, but it also has a number of disadvantages. The majority of the techniques listed below outperform mean imputation.

- Substitution - Impute the value from a new individual who was not selected to be in the sample.
- 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.
- Interpolation and extrapolation - An estimated value from other observations from the same individual. It usually only works in longitudinal data.
- Stochastic regression imputation - The predicted value from a regression plus a random residual value. This has all the advantages of regression imputation but adds in the advantages of the random component.

12. **A/B Testing** - A/B testing data science is a methodical way to evaluate the performance of two variants of a website, app, or campaign. It also goes by the name "split testing." By dividing traffic into two groups and serving one group the A/B version while serving the other group the control, A/B testing seeks to determine what works and doesn't work for your business (the base version). This enables us to evaluate the impact of various versions on conversion rates and response rates.

    For example, you might send two versions of an email to your customer list and figure out which one generates more sales. One version would be the null hypothesis and the other one would be the alternative hypothesis.

    Steps:

- **Make a Hypothesis**
  A hypothesis is an unproven assumption about how the natural world works. However, if it turns out to be accurate, it may help to explain some facts or observations.

- **Null hypothesis or H0:**
  The assumption that sample observations are solely the consequence of chance is known as the null hypothesis. The null hypothesis states that there is no difference between the control and test groups from the perspective of A/B testing data. "There is no difference in the traffic brought by A and B" could be H0 in this case.

- **Alternative Hypothesis or Ha:**
  The null hypothesis is typically opposed by the alternative hypothesis, which questions it.

13. Imputation is not a recommended practice in general due to the following:
    - It just estimating means: mean imputation preserves the mean of the observed data
    - Leads to an underestimation of standard deviation

- Distorts relationships between variables by “pulling” estimates of the correlation toward zero

14. **Linear regression** models the relationships between at least one explanatory variable and an outcome variable. These variables are known as the independent and dependent variables, respectively. When there is one independent variable (IV), the procedure is known as simple linear regression. When there are more IVs, it is referred to as multiple regression.
    Linear regression has two primary purposes—understanding the relationships between variables and forecasting.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

    - The coefficients represent the estimated magnitude and direction (positive/negative) of the relationship between each independent variable and the dependent variable.
    - A linear regression equation allows you to predict the mean value of the dependent variable given values of the independent variables that you specify.

15. **Branches of Statistics**:

    Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

    Descriptive statistics are also categorised into four different categories:
    - Measure of frequency
    - Measure of dispersion
    - Measure of central tendency
    - Measure of position

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information. For example, deriving estimates from hypothetical research.