Course: Educational Statistics (8614)
Semester: Spring, 2023

ASSIGNMENT No. 2

Q.1 How do we calculate median? Also mention its merits and demerits.

The median is a measure of central tendency used in statistics to describe the middle value of a dataset. It is a robust statistic that is less affected by extreme values or outliers compared to the mean. The median divides the dataset into two equal halves, where half the values are greater than or equal to the median and the other half are less than or equal to the median.

To calculate the median, follow these steps:

1. Arrange the dataset in ascending or descending order.
2. If the number of observations is odd, the median is the middle value of the ordered dataset.
3. If the number of observations is even, the median is the average of the two middle values.

Merits of Median:

1. Robustness: The median is less sensitive to outliers or extreme values compared to the mean. It provides a better representation of the central value in skewed distributions.
2. Ease of Interpretation: The median represents the middle value, which is intuitively understandable and doesn't require complex calculations.
3. Applicability to Ordinal Data: The median can be calculated for ordinal data, where the values have a specific order but lack precise numerical meaning.

Demerits of Median:

1. Less Efficient: The median discards some information by only considering the middle values and not the entire dataset. It may not accurately represent the overall distribution of the data.
2. Limited Statistical Properties: The median has fewer statistical properties compared to the mean, making it less suitable for certain mathematical operations or modeling techniques.
3. Potential Ambiguity: In datasets with repeated values or a small range of values, the median may fall between two observations, causing ambiguity in interpretation.

In summary, the median is a useful measure of central tendency, especially when dealing with skewed distributions or outliers. However, it has limitations in terms of efficiency and mathematical properties compared to the mean.

Q.2 Explain the process and errors in hypothesis testing.

Hypothesis testing is a statistical method used to make inferences and draw conclusions about a population based on sample data. It involves setting up two competing hypotheses, the null hypothesis (H0) and the alternative hypothesis (H1), and assessing the evidence against the null hypothesis using sample data.

The process of hypothesis testing can be summarized in the following steps:

Step 1: State the hypotheses:

- Null Hypothesis (H0): It represents the status quo or the assumption to be tested. It assumes no significant difference or relationship between variables.
- Alternative Hypothesis (H1): It contradicts the null hypothesis and suggests that there is a significant difference or relationship between variables.

Step 2: Set the significance level:

- The significance level (often denoted as α) determines the threshold for accepting or rejecting the null hypothesis. Commonly used values for α are 0.05 or 0.01, representing a 5% or 1% chance of rejecting the null hypothesis when it is true.

Step 3: Collect and analyze the data:

- Collect a representative sample from the population of interest.
- Perform appropriate statistical analysis based on the nature of the data and the research question.
- Calculate test statistics that quantify the observed difference or relationship between variables.

Step 4: Determine the critical region and p-value:

- The critical region is the range of values of the test statistic that, if observed, leads to rejection of the null hypothesis.

- The p-value represents the probability of obtaining a test statistic as extreme as or more extreme than the observed value, assuming the null hypothesis is true. A smaller p-value indicates stronger evidence against the null hypothesis.

Step 5: Make a decision and draw conclusions:

If the test statistic falls within the critical region or the p -value is smaller than the chosen significance level (α), the null hypothesis is rejected in favor of the alternative hypothesis. This suggests that there is evidence to support the alternative hypothesis.

- If the test statistic falls outside the critical region or the p-value is greater than the significance level (α), the null hypothesis is not rejected. This indicates that there is insufficient evidence to support the alternative hypothesis.

Errors in Hypothesis Testing:

1. Type I Error (False Positive): This occurs when the null hypothesis is wrongly rejected, indicating a significant difference or relationship when it doesn't exist in the population. The probability of Type I error is equal to the chosen significance level (α). It represents the risk of drawing a false conclusion.
2. Type II Error (False Negative): This occurs when the null hypothesis is wrongly accepted, suggesting no significant difference or relationship when there actually is one in the population. The probability of Type II error is denoted as β. It represents the risk of failing to detect a true effect or relationship.

The balance between Type I and Type II errors depends on factors such as the sample size, effect size, and chosen significance level. It is often denoted as the power of the test, which is the probability of correctly rejecting the null hypothesis when it is false (1 - β).

Hypothesis testing is a fundamental tool in statistical analysis and scientific research. It provides a systematic approach to evaluate the evidence against a null hypothesis and make informed conclusions about populations based on sample data. However, it is important to interpret the results with caution, considering the potential for errors and the limitations of the statistical tests employed.

Q.3 What do you understand by 'Pearson Correlation'? Where is it used and how is it interpreted?

Pearson correlation, also known as Pearson's correlation coefficient or Pearson's r, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It assesses how closely the data points of the two variables align around a straight line.

The Pearson correlation coefficient, denoted by the symbol "r," ranges from -1 to +1. The value of r indicates the degree of association between the variables:

- A positive value of r (between 0 and +1) indicates a positive linear relationship, where as one variable increases, the other tends to increase as well.
- A negative value of r (between -1 and 0) indicates a negative linear relationship, where as one variable increases, the other tends to decrease.
- A value of r close to 0 indicates a weak or no linear relationship between the variables.

The formula to calculate Pearson correlation coefficient is:

$r = (\Sigma((X - \bar{X})(Y - \bar{Y}))) / (\text{sqrt}(\Sigma(X - \bar{X})^2) * \text{sqrt}(\Sigma(Y - \bar{Y})^2))$

where X and Y are the values of the two variables, $\bar{X}$ and $\bar{Y}$ are their respective means, and Σ represents the summation symbol.

Pearson correlation is widely used in various fields, including:

1. Research and Social Sciences: It is used to examine relationships between variables in social science research, such as studying the correlation between education level and income, or analyzing the association between variables in psychology or sociology.
2. Economics and Finance: Pearson correlation helps assess the relationship between economic indicators or financial variables, such as the correlation between stock prices and interest rates, or the relationship between GDP and unemployment rates.
3. Medicine and Health Sciences: It is used to investigate relationships between variables in medical research, such as studying the correlation between blood pressure and body mass index (BMI) or analyzing the association between risk factors and disease outcomes.

4. Data Analysis and Data Mining: Pearson correlation is employed in exploratory data analysis and data mining to identify associations between variables and to screen for potential predictors or factors of interest.

Interpreting the Pearson correlation coefficient involves considering both the magnitude and direction of the correlation:

- If |r| is close to 1, it indicates a strong linear relationship between the variables. The closer r is to +1 or -1, the stronger the relationship.
- If |r| is close to 0, it suggests a weak or no linear relationship between the variables.

It is important to note that Pearson correlation only measures linear relationships and may not capture non-linear associations. Additionally, correlation does not imply causation, and other factors or variables may influence the relationship between the variables being analyzed.

In summary, Pearson correlation is a widely used statistical measure to assess the strength and direction of the linear relationship between two continuous variables. It helps researchers and analysts understand the association between variables and can provide valuable insights in various fields of study and data analysis.

Q.4 Explain ANOVA and its logics.

ANOVA (Analysis of Variance) is a statistical method used to compare the means of two or more groups to determine if there are significant differences among them. It assesses the variation between groups and within groups to make inferences about population means. ANOVA is based on the logic of partitioning the total variation in the data into different sources of variation.

The logic of ANOVA can be explained as follows:

1. Null Hypothesis: The null hypothesis (H0) in ANOVA assumes that there is no significant difference in the means of the groups being compared. It suggests that any observed differences are due to random sampling variability.
2. Alternative Hypothesis: The alternative hypothesis (H1) in ANOVA contradicts the null hypothesis and states that there is a significant difference in at least one of the group means. It implies that the observed differences are not solely due to chance.

3. Variation and Sum of Squares: ANOVA breaks down the total variation in the data into two components: the variation between groups and the variation within groups. This is done by calculating the sum of squares (SS) for each component.

- Sum of Squares Between (SSB): This measures the variation between the group means. It quantifies how much the means differ from each other.
- Sum of Squares Within (SSW): This measures the variation within each group. It quantifies the variability of individual data points within each group.

4. Degrees of Freedom: Degrees of freedom (df) represent the number of independent pieces of information available for estimating the population parameters. In ANOVA, there are two types of degrees of freedom:

- Degrees of Freedom Between (dfB): This is the number of groups minus one.
- Degrees of Freedom Within (dfW): This is the total number of observations minus the number of groups.

5. Mean Squares: Mean squares are calculated by dividing the sum of squares by their respective degrees of freedom:

- Mean Square Between (MSB) = SSB / dfB
- Mean Square Within (MSW) = SSW / dfW

6. F-Statistic: The F-statistic is the ratio of the mean square between to the mean square within:

- F = MSB / MSW

7. F-Test and Decision: The F-statistic is compared to the critical value from the F-distribution based on the chosen significance level (α). If the calculated F-statistic is greater than the critical value, the null hypothesis is rejected, indicating that there are significant differences among the group means. If the calculated F-statistic is not greater than the critical value, the null hypothesis is not rejected.

8. Post hoc Tests: If the overall ANOVA test indicates significant differences among the groups, post hoc tests can be performed to identify which specific groups differ from each other. Common post hoc tests

include Tukey's Honestly Significant Difference (HSD), Bonferroni correction, or pairwise t-tests with adjusted p-values.

ANOVA is a powerful tool for comparing means across multiple groups and is commonly used in experimental studies, social sciences, biology, and many other fields. It allows researchers to determine if observed differences among groups are statistically significant and provides insights into the sources of variation in the data.

Q.5 Explain Chi-Square. Also discuss it as independent test.

Chi-Square (χ^2) is a statistical test used to determine if there is a significant association or relationship between two categorical variables. It compares the observed frequencies of categories in a contingency table with the frequencies that would be expected if the variables were independent.

The logic behind the Chi-Square test can be explained as follows:

1. Null Hypothesis: The null hypothesis (H0) in the Chi-Square test assumes that there is no association between the two categorical variables. It suggests that any observed differences are due to random chance or sampling variability.
2. Alternative Hypothesis: The alternative hypothesis (H1) contradicts the null hypothesis and states that there is a significant association between the two categorical variables. It implies that the observed differences are not solely due to chance.
3. Contingency Table: The Chi-Square test uses a contingency table to organize the observed frequencies of the categories for each variable. The table displays the joint distribution of the two variables.
4. Expected Frequencies: The Chi-Square test calculates the expected frequencies for each cell in the contingency table under the assumption of independence. The expected frequency is calculated by multiplying the row total and column total and dividing by the grand total.
5. Chi-Square Statistic: The Chi-Square test statistic is calculated by comparing the observed frequencies with the expected frequencies in each cell of the contingency table. The formula for calculating the Chi-Square statistic is:

   $\chi^2 = \Sigma((O - E)^2 / E)$

where O is the observed frequency and E is the expected frequency for each cell.

6. Degrees of Freedom: Degrees of freedom (df) in the Chi-Square test depend on the dimensions of the contingency table. For a 2x2 table, the degrees of freedom is 1. For larger tables, the degrees of freedom are calculated as (r - 1) x (c - 1), where r is the number of rows and c is the number of columns in the table.
7. Chi-Square Distribution: The Chi-Square test statistic follows a Chi-Square distribution with degrees of freedom equal to the number of degrees of freedom calculated in the previous step.
8. Critical Value and p-value: The Chi-Square test compares the calculated Chi-Square statistic with the critical value from the Chi-Square distribution based on the chosen significance level (α). If the calculated Chi-Square statistic is greater than the critical value, the null hypothesis is rejected, indicating a significant association between the variables. The p-value associated with the Chi-Square statistic can also be calculated, which represents the probability of obtaining a Chi-Square value as extreme as or more extreme than the observed value, assuming the null hypothesis is true.

Chi-Square as an Independent Test: Chi-Square test is commonly used as an independent test when there are two categorical variables and the goal is to determine if there is an association between them. It can be used to analyze survey responses, examine the relationship between two categorical variables in a population, assess the effectiveness of a treatment in different groups, and more.

The Chi-Square test as an independent test allows us to draw conclusions about the relationship between variables based on the observed frequencies and expected frequencies in the contingency table. By comparing the calculated Chi-Square statistic with the critical value or p-value, we can make decisions regarding the presence or absence of an association.

It is important to note that the Chi-Square test assesses the association between categorical variables but does not provide information about the strength or direction of the relationship. Additional measures such as Cramér's V or contingency coefficients can be used to quantify the strength of the association.

In summary, the Chi-Square test is a statistical test used to determine the presence of a significant association between two categorical variables. It compares the observed frequencies in a contingency table with the

expected frequencies under the assumption of independence. The test involves formulating null and alternative hypotheses, calculating the Chi-Square test statistic, determining degrees of freedom, and comparing the test statistic with critical values or p-values.

As an independent test, the Chi-Square test helps researchers analyze and interpret data to understand the relationship between categorical variables. It is widely used in social sciences, market research, epidemiology, and other fields where categorical data analysis is essential. By conducting the Chi-Square test, researchers can gain insights into the presence or absence of an association, which can inform decision-making and further research.