Statistical Methods

Md. Jubayer Hossain

2023-06-12

Median (IQR); n (%)

"Median (IQR); n (%)" is a statistical notation commonly used to present descriptive statistics for a categorical variable or a subgroup within a dataset. It combines information about the median, the interquartile range (IQR), and the frequency or proportion of observations in each category.

1. Median (IQR): The median represents the middle value of a dataset, while the IQR indicates the range of values that span the middle 50% of the data. The IQR is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). Typically, the median and IQR are provided together in parentheses to summarize the central tendency and spread of a continuous variable.

2. n (%): The "n" represents the frequency or count of observations in a particular category or subgroup. It indicates how many data points fall into that specific group. The percentage (%) represents the proportion of observations in each category relative to the total sample size.

By combining these two components, "n (%)" provides a summary of both the frequency and relative distribution of observations within each category or subgroup, allowing for a comprehensive understanding of the dataset. This notation is commonly used in tables or summaries when reporting descriptive statistics for categorical variables, allowing researchers or readers to quickly grasp the key features of the data.

Wilcoxon rank sum test

The Wilcoxon rank sum test, also known as the Mann-Whitney U test, is a nonparametric statistical test used to assess whether there is a significant difference between the distributions of two independent groups. It is often employed when the assumptions of normality and equal variances required by parametric tests, such as the t-test, are not met.

- H_0: There is no significant difference between the distributions of the two groups.
- H_a: There is a significant difference between the distributions of the two groups.

Pearson's Chi-squared test

Pearson's chi-squared test, also known as the chi-square test, is a statistical test used to determine whether there is a significant association between two categorical variables. It compares the observed frequencies in a contingency table to the frequencies that would be expected if the variables were independent.

- H_0: There is no significant association between the two categorical variables.
- H_a: There is a significant association between the two categorical variables.

Fisher's exact test

Fisher's exact test is a statistical test used to determine the significance of the association between two categorical variables in a 2x2 contingency table when the sample sizes are small. It is a nonparametric test and is commonly employed when the assumptions for the chi-square test, such as expected cell frequencies being greater than 5, are not met.

- H_0: There is no significant association between the two categorical variables.
- H_a: There is a significant association between the two categorical variables.

Logistic Regression

Logistic regression is a statistical modeling technique used to predict binary or categorical outcomes based on a set of predictor variables. The coefficients obtained from logistic regression can be exponentiated to calculate odds ratios (ORs), which provide insights into the relationships between the predictor variables and the probability of the outcome occurring.

Here's how to interpret odds ratios in logistic regression: 1. An odds ratio (OR) greater than 1 indicates that there is a positive association between the predictor variable and the outcome variable. For example, an OR of 1.5 suggests that for each unit increase in the predictor variable, the odds of the outcome variable occurring are 1.5 times higher.

2. An odds ratio (OR) less than 1 indicates that there is a negative association between the predictor variable and the outcome variable. For example, an OR of 0.7 suggests that for each unit increase in the predictor variable, the odds of the outcome variable occurring are 0.7 times lower.

3. An odds ratio (OR) equal to 1 suggests no association between the predictor variable and the outcome variable. In other words, the predictor variable does not influence the odds of the outcome occurring.

4. If the OR is greater than 1 and statistically significant, it implies that the predictor variable has a significant positive effect on the odds of the outcome variable occurring.

5. If the OR is less than 1 and statistically significant, it implies that the predictor variable has a significant negative effect on the odds of the outcome variable occurring.

It is important to consider the confidence interval (CI) associated with the odds ratio. If the CI does not include 1, the odds ratio is considered statistically significant at a chosen significance level (e.g., 0.05). A CI entirely above 1 suggests a significant positive effect, while a CI entirely below 1 suggests a significant negative effect.

Note that the interpretation of odds ratios assumes that the other predictor variables in the logistic regression model are held constant. Additionally, the interpretation may vary depending on the coding scheme and scaling of the predictor variables. It's always important to carefully interpret the odds ratios in the specific context of the study and the variables involved.

P-value Interpretations

When performing hypothesis tests, the p-value is a measure of the strength of evidence against the null hypothesis. It quantifies the probability of observing the data or more extreme results under the assumption that the null hypothesis is true. In general, a smaller p-value indicates stronger evidence against the null hypothesis.

When interpreting the p-value, a commonly used significance level is 0.05, which corresponds to a 5% chance of observing the data or more extreme results if the null hypothesis is true.

If the obtained p-value is less than 0.05 (the chosen significance level), it suggests that the observed data is unlikely to have occurred by chance alone, assuming the null hypothesis is true. In this case, we reject the null hypothesis and conclude that there is sufficient evidence to support the alternative hypothesis.

On the other hand, if the p-value is greater than or equal to 0.05, we do not have enough evidence to reject the null hypothesis. This means that the observed data is reasonably likely to occur by chance alone, assuming the null hypothesis is true.

Regarding the interpretation of a 95% confidence interval (CI), it provides a range of plausible values for a population parameter. If the interval does not include the null value (e.g., zero for the difference between two groups), it suggests that the parameter is significantly different from the null value at the 5% significance level.

In summary, when the p-value is less than 0.05, it indicates statistical significance, suggesting evidence against the null hypothesis. When the 95% confidence interval does not include the null value, it also supports rejecting the null hypothesis.