How to Calculate the Correlation Between Nominal Input Features and Ordinal Output

When you're analyzing data, understanding the relationship between your input features and the output variable is a crucial step. This becomes a bit more complex when your input features are **nominal** and your output is **ordinal**.

In simple terms:

- **Nominal input**: This refers to categorical data where there is no inherent order between the categories. Examples include variables like gender (male, female), region (north, south, east, west), or product type (A, B, C).
- **Ordinal output**: This refers to variables where the categories have a meaningful order but the distances between them are not uniform. Examples include educational levels (high school, bachelor's, master's), customer satisfaction ratings (poor, average, excellent), or performance levels (low, medium, high).

The goal here is to determine how the nominal input variable (which is categorical without a specific order) influences or relates to the ordinal output (which has an inherent order but unequal intervals). In this blog, we'll explore methods to calculate this relationship and give you the tools to analyze these kinds of data effectively.

Why is This Problem Interesting?

Nominal variables, being categorical, don't have an order, while ordinal variables do. The challenge in analyzing their relationship lies in how we deal with the fact that the nominal input does not have inherent order or ranking, while the ordinal output does. Despite the differences in their nature, there are several statistical methods that can help uncover associations between these types of variables.

Let's dive into the most common techniques used to analyze the correlation between **nominal input features** and **ordinal output variables**.

---

1. Chi-Square Test of Independence

The **Chi-Square Test of Independence** is one of the most commonly used tests for examining the relationship between categorical variables. Although the Chi-square test doesn't directly

calculate correlation in the sense of measuring strength or direction, it is often used to assess whether there is an association between the nominal input and the ordinal output.

How It Works:

1. **Create a contingency table**: This table will show the frequencies of the different combinations of your nominal input categories and ordinal output categories.
2. **Calculate the Chi-square statistic**: This statistic compares the observed frequencies (i.e., how often each combination of nominal and ordinal categories occurs) to the expected frequencies (i.e., what we would expect if the two variables were independent).
3. **Interpret the result**: The p-value from the Chi-square test indicates whether the relationship between the nominal input and ordinal output is statistically significant.
    - A **low p-value (usually < 0.05)** suggests that the nominal input and ordinal output are **not independent** and that there is a significant relationship.
    - A **high p-value** indicates that the variables are **independent**.

Why Use It:

- The Chi-square test works well when you want to check for associations between nominal and ordinal variables.
- It is simple and can be used when you have categorical data, although it doesn't provide a measure of the strength or direction of the relationship.

Example:

If you're studying customer satisfaction (ordinal: "poor," "fair," "good") and region (nominal: "North," "South," "East," "West"), the Chi-square test will help determine if there's an association between customer satisfaction levels and the region.

---

2. Ordinal Logistic Regression

Ordinal Logistic Regression, also known as **proportional odds regression**, is a modeling approach designed for situations where the dependent variable is ordinal, and it's used to understand the relationship between one or more predictors (which can be nominal, ordinal, or continuous) and an ordinal outcome.

How It Works:

- Ordinal logistic regression models the probability that an observation falls into a particular category of the ordinal output (or higher) based on the values of the nominal input.
- In this case, the **nominal input** would be treated as a categorical predictor in the model, and the **ordinal output** would be modeled as an ordered categorical variable.

Why Use It:

- Ordinal logistic regression is particularly useful if you want to model the probability of different levels of the ordinal output based on the nominal input.
- It's ideal when you want to predict or assess the effect of nominal input features on an ordinal outcome.

Example:

If you are looking at the relationship between a person's **region** (nominal: North, South, East, West) and their **job satisfaction** level (ordinal: low, medium, high), ordinal logistic regression can help predict how likely it is that someone from a specific region will report a particular level of job satisfaction.

---

3. Kruskal-Wallis H Test

The **Kruskal-Wallis H Test** is a non-parametric method used to test if there are statistically significant differences between the distributions of the ordinal output variable across the different categories of the nominal input variable.

How It Works:

1. **Group the data**: The Kruskal-Wallis test groups the data based on the categories of the nominal variable (e.g., different regions).
2. **Compare the ranks**: It then compares the ranks of the ordinal output variable within each group.
3. **Test for differences**: The test determines whether the ranks of the ordinal output are similar across the groups defined by the nominal input.
    - A **low p-value (typically < 0.05)** suggests that there are significant differences in the ordinal output based on the categories of the nominal input.
    - A **high p-value** means there are no significant differences between the groups.

Why Use It:

- The Kruskal-Wallis test is useful when you want to assess whether there are differences in the ordinal output variable between multiple categories of the nominal input.
- Unlike analysis of variance (ANOVA), which assumes normally distributed data, Kruskal-Wallis is non-parametric, making it more flexible for data that doesn't follow a normal distribution.

Example:

If you want to see whether employees from different **regions** (nominal) report different levels of **job satisfaction** (ordinal: "low," "medium," "high"), the Kruskal-Wallis H test would help you understand if there are significant differences in job satisfaction across regions.

---

4. Cramér's V (for Chi-Square Results)

Once you've performed the **Chi-Square Test of Independence**, you can also calculate **Cramér's V**, a measure of association for nominal and ordinal data. Cramér's V ranges from 0 to 1 and tells you how strong the association is between your nominal input and ordinal output.

How It Works:

Cramér's V is calculated using the Chi-square statistic and the sample size. The formula is:

V=χ2n(k−1)V = \sqrt{\frac{\chi^2}{n(k - 1)}}

Where:

- χ2\chi^2 is the Chi-square statistic,
- nn is the total number of observations,
- kk is the smaller number of categories between the nominal input and ordinal output variables.

Why Use It:

- Cramér's V provides a **quantitative measure** of the strength of the relationship between the nominal and ordinal variables. It's particularly useful when you want to go beyond a simple Chi-square test and quantify how strong the association is.

Example:

After performing the Chi-square test between **region** (nominal) and **job satisfaction** (ordinal), you can compute Cramér's V to determine how strong the relationship is, with values closer to 1 indicating a strong association.

Summary of Methods

Method	Best Used For	Strength
Chi-Square Test of Independence	Testing if there is an association between nominal input and ordinal output	Simple and widely used, but doesn't give a measure of strength or direction

Ordinal Logistic Regression	Modeling the probability of different levels of the ordinal output based on nominal input	Predictive and provides insights into the effect of the nominal input on the ordinal output
Kruskal-Wallis H Test	Testing for differences in the ordinal output across different categories of the nominal input	Non-parametric and flexible for non-normally distributed data
Cramér's V	Quantifying the strength of the association after performing a Chi-square test	Provides a numerical measure of the strength of the association

---

Conclusion

When your **input** is **nominal** and your **output** is **ordinal**, you're dealing with a mix of categorical and ordered variables. The methods outlined in this blog—**Chi-square test**, **Ordinal Logistic Regression**, **Kruskal-Wallis H Test**, and **Cramér's V**—are the key tools you can use to assess the relationship between the two.

- The **Chi-square test** and **Cramér's V** give you insight into whether there's a relationship and how strong it is.
- **Ordinal Logistic Regression** is perfect for predictive models, especially when you want to model probabilities for different ordinal output categories.
- The **Kruskal-Wallis H Test** allows you to test for significant differences in the ordinal output across different categories of the nominal input.

By selecting the right method based on your data, you can gain a deeper understanding of how nominal input features relate to ordinal output variables and make more informed decisions in your analysis.