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

In data analysis, one of the most interesting relationships you can explore is between two **nominal variables**. Nominal variables represent categories without any intrinsic order—think of variables like "color," "product type," "city name," or "gender." When both your **input feature** and **output variable** are **nominal**, you're trying to understand how changes in one categorical variable might be associated with changes in another categorical variable. For example, you might want to explore how **marketing campaigns** (nominal: "TV," "Email," "Social Media") relate to **purchase decisions** (nominal: "Yes," "No").

So, how do you measure the relationship, or **correlation**, between two nominal variables? In this blog post, we'll discuss the methods and techniques for calculating correlation in this scenario, where both the input and output are categorical in nature.

What Are Nominal Variables?

Before diving into the methods, let's make sure we understand **nominal variables**:

- **Nominal input**: These are variables that represent categories with no inherent order or ranking. Examples include:
    - **Color**: "Red," "Blue," "Green"
    - **Region**: "North," "South," "East," "West"
    - **Product Type**: "A," "B," "C"
- **Nominal output**: Similarly, these are categorical variables with no order. Examples could include:
    - **Purchase Decision**: "Yes," "No"
    - **Customer Satisfaction**: "Low," "Medium," "High"
    - **Region of Preference**: "Urban," "Suburban," "Rural"

When both the input and output variables are nominal, the goal is to identify any association or pattern between them, like whether certain categories of the input variable are more likely to lead to certain outcomes in the output variable.

---

1. Chi-Square Test of Independence

The most common method for calculating the correlation between two nominal variables is the **Chi-Square Test of Independence**. This test evaluates whether two categorical variables are **independent** of each other or whether there is an association between them.

How It Works:

The Chi-Square test compares the **observed frequencies** of different combinations of categories in a contingency table to the **expected frequencies** if the two variables were independent. If the difference between observed and expected values is large, it suggests that there is an association between the variables.

The test statistic for Chi-Square is given by:

χ2=∑(Oi−Ei)2Ei\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}

Where:

- OiO_i is the observed frequency for category ii,
- EiE_i is the expected frequency for category ii,
- The summation is over all possible categories.

Why Use It:

The Chi-Square test is ideal when both the input and output variables are nominal. It helps determine whether there's a significant association between the two categorical variables.

Example:

Let's say you're studying the relationship between **product type** (nominal: "A," "B," "C") and **customer satisfaction** (nominal: "Low," "Medium," "High"). You can create a contingency table to represent the frequencies of each combination of product type and satisfaction level, and then perform the Chi-Square test to see if satisfaction is independent of product type, or if there's a significant association between them.

Steps in Chi-Square:

1. **Create a contingency table**: This table summarizes the observed counts of each combination of the categories in the two variables.
2. **Calculate expected frequencies**: These are the frequencies you would expect if there was no association between the variables.
3. **Compute the Chi-Square statistic**: Compare observed and expected frequencies.
4. **Interpret the p-value**: A small p-value (typically < 0.05) suggests that the variables are **not independent**, indicating a significant relationship between the two nominal variables.

---

2. Cramér's V

While the Chi-Square test tells you whether there is an association between the variables, it doesn't quantify the strength of that association. For this, you can use **Cramér's V**, a measure of association based on the Chi-Square statistic.

How It Works:

Cramér's V is calculated from the Chi-Square statistic and the sample size. It normalizes the Chi-Square value, so it provides a measure of association that is not affected by the size of the table.

The formula for Cramér's V is:

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

Where:

- $\chi 2\chi^2$ is the Chi-Square statistic,
- nn is the total sample size,
- kk is the smaller of the number of rows or columns in the contingency table.

Cramér's V gives a value between 0 and 1:

- **0** indicates no association between the variables,
- **1** indicates a perfect association.

Why Use It:

Cramér's V is useful when you want to measure the **strength** of the association between two nominal variables. It provides a clearer picture of how strongly the variables are related.

Example:

After performing a Chi-Square test to check the association between **product type** and **customer satisfaction**, you can calculate Cramér's V to quantify the strength of this relationship. A Cramér's V value closer to 1 would indicate a strong association, while a value closer to 0 suggests a weak or no association.

---

3. Contingency Table and Visual Analysis

While statistical tests like Chi-Square and Cramér's V are essential for quantifying the relationship between two nominal variables, it's often useful to start with a **contingency table**

and visually examine the data. A contingency table displays the frequency of each combination of the categories from the two variables.

How It Works:

A contingency table is essentially a matrix where the rows represent the categories of the input variable, and the columns represent the categories of the output variable. The entries in the table show the counts of occurrences for each combination of input and output categories.

Why Use It:

Contingency tables provide a clear, visual understanding of the distribution of data across the different categories. It can help identify patterns or trends that may not be immediately obvious.

Example:

You might create a contingency table for **product type** and **customer satisfaction**, showing how many customers who bought each product type were satisfied at each level. This visualization can give you an intuitive sense of any associations between the variables, even before applying statistical tests.

---

4. Logistic Regression (for Binary Nominal Output)

If the output variable is binary (e.g., "Yes" or "No"), you can use **logistic regression** to model the relationship between a nominal input and the binary output. While logistic regression is typically used for prediction, it also provides insight into the association between categorical variables.

How It Works:

Logistic regression estimates the probability of an event occurring based on the input features. It can handle nominal input variables by using techniques like **dummy coding** (where each category of the nominal input is represented by a binary variable).

Why Use It:

Logistic regression is useful if you’re trying to predict a binary outcome based on one or more nominal inputs and want to quantify how changes in the input affect the likelihood of the output.

Example:

If you want to understand how different **marketing strategies** ("Email," "Social Media," "TV") influence whether a customer **purchases a product** ("Yes" or "No"), logistic regression can provide insights into the strength and direction of the relationship.

Summary of Methods

Method	Best Used For	Strength
Chi-Square Test of Independence	Testing if two nominal variables are independent or associated	Simple and widely used for categorical variables
Cramér's V	Measuring the strength of the association between two nominal variables	Normalizes Chi-Square for easier interpretation of association strength
Contingency Table	Visualizing the distribution of data between two categorical variables	Helps identify patterns and trends visually
Logistic Regression (Binary Output)	Predicting binary outcomes based on nominal input variables	Provides a way to quantify how nominal inputs influence binary outcomes

---

Conclusion

When both the **input feature** and the **output variable** are **nominal**, several methods can help you measure the relationship or association between the two. Whether you're performing a **Chi-Square test of independence** to see if there's a significant association, calculating **Cramér's V** to assess the strength of the relationship, or visualizing the data with a **contingency table**, each method has its place in helping you understand the connection between categorical variables.

By applying the appropriate statistical techniques, you can gain valuable insights into how different categories of your nominal input variable influence or are associated with your nominal output variable, which can help guide decision-making, marketing strategies, and more.