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

In the world of data analysis, understanding the relationship between variables is crucial for building effective models and extracting insights from data. When both the input features and the output variable are **ordinal**, meaning that they have a meaningful order or ranking but no consistent interval between the categories, determining the correlation between them is slightly different from traditional correlation methods.

In this blog post, we'll explore how to calculate the correlation between **ordinal input features** (e.g., levels of customer satisfaction: "poor," "good," "excellent") and **ordinal output variables** (e.g., employee performance ratings: "low," "medium," "high").

We'll go over the methods used to assess this relationship, how they work, and which one you should choose based on your data and analysis goals.

What is Ordinal Data?

Before diving into the specifics of calculating correlation, let's quickly review **ordinal data**. Ordinal data refers to variables that have a clear ordering or ranking of categories, but the distances between these categories are not consistent. For example:

- **Ordinal Input (e.g., Customer Satisfaction)**: "poor," "average," "good," "excellent."
- **Ordinal Output (e.g., Employee Performance)**: "low," "medium," "high."

The key feature of ordinal data is that the categories have an inherent order, but we don't know the exact difference between each category (i.e., the intervals are not uniform). Because of this, methods used to calculate correlation for ordinal data must respect the ordered nature of the variables.

Methods to Calculate Correlation Between Ordinal Input and Ordinal Output

When both the input and the output variables are ordinal, you need specialized methods that account for the ordering of the categories. Below are the most common techniques to calculate the correlation between ordinal variables.

---

1. Spearman's Rank Correlation

Spearman's rank correlation is one of the most popular and widely used methods for calculating the relationship between two ordinal variables. It measures how well the ranks of the ordinal input feature correspond to the ranks of the ordinal output. Unlike Pearson's correlation, which assumes a linear relationship, Spearman's rank correlation looks for a **monotonic** relationship, meaning that as one variable increases, the other tends to increase (or decrease) consistently, but not necessarily at a constant rate.

How It Works:

- **Step 1**: Rank both the ordinal input and the ordinal output categories. For example, if your ordinal input is "poor," "average," and "good," you would assign ranks (1 for "poor," 2 for "average," and 3 for "good").

- **Step 2**: Calculate the difference between the ranks of the two variables for each data point.

- **Step 3**: Square the differences between ranks, sum them up, and plug the sum into the Spearman's formula:

ρ=1−6∑di2n(n2−1)\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}

Where:

- did_i is the difference in ranks for each pair of data points,
- nn is the number of data points.

Why Use It:

Spearman's rank correlation works well because it doesn't require the data to be normally distributed, and it handles the ordinal nature of the variables. It will tell you whether there is a monotonic relationship (i.e., a consistent increase or decrease) between the ordinal input and output.

Example:

Suppose you want to examine whether there's a relationship between customer satisfaction (ordinal) and product rating (ordinal). After ranking both variables, Spearman's rank correlation will calculate a value between -1 and +1:

- **+1** indicates a perfect positive monotonic relationship (as satisfaction increases, the rating increases),
- **-1** indicates a perfect negative monotonic relationship (as satisfaction increases, the rating decreases),
- **0** indicates no relationship.

---

2. Kendall's Tau

Kendall's Tau is another rank-based correlation measure, similar to Spearman's rank, but it's often preferred when you have smaller datasets or when there are many tied ranks (i.e., cases where the ranks are identical). Kendall's Tau measures the strength and direction of association between two ordinal variables by counting the number of concordant and discordant pairs in the data.

How It Works:

- **Concordant pairs** are pairs where the ranks of both variables agree (e.g., if the rank of the input increases, the rank of the output also increases).
- **Discordant pairs** are pairs where the ranks of the two variables disagree (e.g., if the rank of the input increases, the rank of the output decreases).

The formula for Kendall's Tau is:

τ=(C−D)(C+D+T1)(C+D+T2)\tau = \frac{(C - D)}{ \sqrt{(C + D + T_1)(C + D + T_2)}}

Where:

- CC is the number of concordant pairs,
- DD is the number of discordant pairs,
- T1T_1 and T2T_2 are the number of tied ranks in the input and output variables, respectively.

Why Use It:

Kendall's Tau is less sensitive to ties and is often considered more robust than Spearman's rank correlation, especially when dealing with small datasets or datasets with many ties.

Example:

Let's say you are studying the relationship between customer satisfaction (ordinal) and employee performance (ordinal). Kendall's Tau would calculate a coefficient between -1 and +1:

- **+1** indicates a perfect positive relationship (as satisfaction increases, performance also increases),
- **-1** indicates a perfect negative relationship (as satisfaction increases, performance decreases),
- **0** indicates no relationship.

---

3. Ordinal Logistic Regression

Ordinal logistic regression (also known as **proportional odds regression**) is a statistical model that is used to predict an ordinal dependent variable based on one or more independent variables. Unlike Spearman's and Kendall's correlations, ordinal logistic regression allows for a more detailed analysis by modeling the probabilities of each category of the ordinal output based on the ordinal input.

How It Works:

- Ordinal logistic regression models the relationship between the ordinal input and ordinal output by estimating the **odds** that the output will fall into a higher or lower category based on the input.
- It treats the ordinal output as a categorical variable with ordered levels, and it models the probability that the output will fall into each category relative to the others.

The model output will give you **odds ratios**, which describe how the likelihood of the output changing from one category to another is influenced by the ordinal input.

Why Use It:

Ordinal logistic regression is useful when you want to **predict** an ordinal output based on an ordinal input. It can provide deeper insights into the relationship between the variables by giving you not just a correlation coefficient but also probabilistic estimates for each category of the ordinal output.

Example:

Suppose you are analyzing employee satisfaction (ordinal) as the input feature and employee retention status (ordinal: "low," "medium," "high") as the output. Ordinal logistic regression can tell you how each level of employee satisfaction influences the likelihood of different retention levels.

---

4. Somers' D

Somers' D is a measure of association specifically designed for ordinal variables. It quantifies the strength and direction of the relationship between two ordinal variables, taking into account both concordant and discordant pairs. Somers' D is closely related to Kendall’s Tau but is often used when you have an ordinal independent variable and a dependent ordinal outcome.

Why Use It:

Somers' D is particularly useful when you are modeling an ordinal input variable (e.g., levels of education) and an ordinal output variable (e.g., job performance), as it helps quantify the direction and strength of the relationship between the two ordinal variables.

Summary of Correlation Methods

Method	Best Used For	Interpretation
Spearman's Rank	Monotonic relationship between ordinal variables	Correlation coefficient between -1 and +1, with 0 indicating no relationship
Kendall's Tau	Small datasets or datasets with many tied ranks	Correlation coefficient between -1 and +1, with 0 indicating no relationship
Ordinal Logistic Regression	Predicting an ordinal outcome based on an ordinal input	Odds ratios and probability estimates for each category of the ordinal output
Somers' D	Assessing strength of the relationship between ordinal variables	A measure of association between -1 and +1, with 0 indicating no relationship

---

Conclusion

When both the **input** and **output** are **ordinal**, calculating correlation requires methods that respect the inherent ordering of the categories. Techniques such as **Spearman's rank correlation**, **Kendall's Tau**, and **Somers' D** are excellent for measuring the strength and direction of monotonic relationships between two ordinal variables.

For more complex analyses, such as when you need to predict an ordinal outcome, **ordinal logistic regression** can be an effective method. By choosing the right correlation method, you can gain valuable insights into how ordinal variables interact with each other, whether you're looking for simple associations or more predictive relationships.