Lecture Note 6: Simpson's Paradox and Why Correlation is Not Causation

1. Correlation Is Not Causation

People sometimes mistake correlation for causation because they observe a relationship between two variables and assume that one variable causes the other. However, correlation merely indicates that two variables are associated or tend to vary together, without necessarily implying a cause-and-effect relationship. Other factors, not accounted for, may influence both variables or there may be a third variable causing the observed correlation.

More formally, correlation is a statistical measure that quantifies the extent of association between two variables. Previously, we demonstrated that a positive (negative) correlation between X and Y indicates that when X is above its mean, Y tends to be above (below) its mean. Similarly, when Y is above its mean, X tends to be above (below) its mean. Hence, while correlation indicates that changes in one variable are linked to changes in another, it does not necessarily imply a causal relationship.

Below are several reasons why correlation should not be mistaken for causation:

1. Coincidental or Spurious Correlation: Sometimes, variables may appear to be correlated by chance or due to random fluctuations in data, leading to a false impression of causation. For example, there might be a strong correlation between ice cream sales and drowning deaths during the summer months. However, this correlation is coincidental because both variables increase during the summer but are not causally related.

2. Bidirectional and Reverse Causality: Correlation does not distinguish which variable is the cause and which is the effect. It's possible that both variables influence each other, or the apparent cause might be the result of the effect. Consider the correlation between education level and income. While higher education often leads to higher income, it's also true that higher income can afford better education, leading to a bidirectional relationship.

3. Confounding Variables: Other variables not considered in the analysis may influence both the variables being studied, creating a misleading correlation. Suppose there's a correlation between the consumption of ice cream and sunglasses sales. However, the confounding variable here could be the weather, as both variables increase during sunny days.

4. Lack of Temporal Precedence: Correlation does not establish the temporal sequence of events. Just because two variables are correlated at a certain point in time doesn't mean one caused the other. If there's a correlation between a person's weight gain and the onset of depression, it doesn't mean that weight gain caused depression. Depression may have caused changes in eating habits, leading to weight gain.

5. Complex Interactions: Correlation doesn't account for complex relationships or interactions between variables, which may influence the observed correlation without implying causation. In the relationship between exercise and weight loss, there might be a correlation, but it's not

straightforward. Factors like diet, metabolism, and genetics can interact in complex ways, influencing weight loss independently of exercise.

6. Sampling Error: Correlation may arise by chance due to random sampling variation, especially in small or biased samples, leading to incorrect assumptions about causation. Suppose a small survey shows a correlation between drinking coffee and living longer. However, this might be due to chance or biased sampling, rather than a causal relationship.

7. Experimental Design and Control Issues: Correlation studies are often observational and lack the control necessary to establish causation. Controlled experiments are required to isolate the effects of variables and establish causality. If a study finds a correlation between a new medication and improved health outcomes, it doesn't prove causation. Without a controlled experiment where participants are randomly assigned to receive the medication or a placebo, other factors could be influencing the results.

In summary, while correlation measures the degree of association between two variables, it does not imply causation. Several factors contribute to this misconception: coincidental correlation, bidirectional causality, confounding variables, lack of temporal precedence, complex interactions, sampling error, and experimental design issues. Understanding these factors is crucial to avoid mistaking correlation for causation in data analysis and interpretation.

2. Simpson's Paradox

Simpson's paradox is a phenomenon in statistics where a trend appears in different groups of data but disappears or reverses when the groups are combined. In other words, the association observed in each subgroup is reversed when the subgroups are aggregated. This paradox highlights the importance of considering the effects of lurking variables or confounders that may influence the observed relationships between variables.

2.1 Example

In this example, I devised a hypothetical scenario to explore the potential impact of ventilators on reducing COVID-related fatalities among hospitalized patients. It's important to note that both the example and the data presented are entirely fabricated and do not represent the actual efficacy of ventilators.

Imagine the study were undertaken in an area with several hospitals, each maintaining records for their COVID in-patients. These 100 patients were either administered a ventilator or not. Below are the summarized data tables we've collected.

Table: Frequency and Relative Frequency

	No Ventilator	**Ventilator**	**sum**
Survival	60 (30%)	40 (20%)	100 (50%)
Death	40 (20%)	60 (30%)	100 (50%)
sum	100 (50%)	100 (50%)	200

Table: Conditional Probabilities

	No Ventilator	**Ventilator**	**sum**
Given Survival	60/100 = 60%	40/100 = 40%	100 (100%)
Given Death	40/100 = 40%	60/100 = 60%	100 (100%)

	Given No Ventilator	**Given Ventilator**
Survival	60/100 = 60%	40/100 = 40%
Death	40/100 = 40%	60/100 = 60%
sum	100 (100%)	100 (100%)

The table illustrates that among surviving patients, 60% did not receive a ventilator, while 40% did. Conversely, among deceased patients, 40% did not receive a ventilator, while 60% did. Similarly, among patients without a ventilator, 60% survived, whereas among patients with a ventilator, only 40% survived. Can we infer that the ventilator has a negative effect on the survival rate of COVID patients?

The answer is negative. What might have gone wrong? Here's one potential explanation: COVID patients admitted to hospitals may have had either severe or extremely severe symptoms. Doctors are more inclined to assign a ventilator to patients with extremely severe symptoms. The availability of ventilators varied among hospitals, and some patients with severe symptoms received a ventilator when one was available at their respective hospitals.

Suppose we've also gathered data on the severity of symptoms for those patients. The tables below offer a more detailed breakdown of the number of patients receiving or not receiving a ventilator based on the severity of their symptoms.

Table: Ventilator Assignment Breakdown

	Given Severe	**Given Extremely Severe**	**sum**
No Ventilator	90 (90%)	10 (10%)	100
Ventilator	10 (10%)	90 (90%)	100
sum	100	100	200

Table: Further breakdown - No Ventilator Patients

	No Ventilator	**Ventilator**	**sum**
Survival	60	0	60
Death	30	10	40
sum	90	10	100

Table: Further breakdown - Ventilator Patients

	No Ventilator	Ventilator	sum
Survival	10	30	40
Death	0	60	60
sum	10	90	100

Now, let's apply our understanding of conditional probability to this scenario, focusing solely on the relationship between ventilator use and survival.

- Probability of Survival given Ventilator: 40%
- Probability of Survival given No Ventilator: 60%

When considering the decision to assign ventilators:

- Among patients with severe symptoms:
  - Probability of Survival given Ventilator: 100%
  - Probability of Survival given No Ventilator: 66.7%
- Among patients with extremely severe symptoms:
  - Probability of Survival given Ventilator: 33.3%
  - Probability of Survival given No Ventilator: 0%

With this additional insight into how ventilators were allocated based on symptom severity, it becomes evident that ventilators contribute to improved survival rates. Previously, we observed lower survival rates among ventilated patients, potentially due to a higher proportion of patients with extremely severe symptoms receiving ventilator support, who naturally had lower survival rates.

2.2 Underlying Causes of Simpson's Paradox

- Confounding Variables:
  Simpson’s Paradox frequently occurs when confounding variables are present, as clearly demonstrated in the example provided above. These third variables, also referred to as confounders, lurking variables, or covariates, exert an influence on both the variables under study. This influence can distort the apparent relationship between the variables being considered, leading to erroneous conclusions. Identifying and accounting for confounding variables is crucial for unraveling the true association between the variables of interest.

- Sample Size Bias:
  Another common cause of Simpson’s Paradox is sample size bias. When different subgroups within the data have vastly different sample sizes, the aggregated data may exhibit trends that are skewed by the unequal distribution of observations. As a result, the apparent relationship between variables can be misleading, particularly when interpreting aggregated statistics. Addressing sample size imbalances is essential for obtaining accurate insights from the data.

- Non-linear Relationships:

Simpson’s Paradox can also arise when the variables involved exhibit non-linear relationships. In such cases, the direction or strength of the relationship between the variables may vary across different subsets of the data. When aggregated, these non-linear patterns can obscure the underlying dynamics, leading to erroneous interpretations. Recognizing the presence of non-linear relationships and accounting for them appropriately is essential for uncovering the true nature of the associations between variables.

2.3 Combatting Confounders

A critical strategy for addressing confounding variables is employing a technique known as stratification. Stratifying the data involves dividing the dataset into distinct subgroups or strata based on specific characteristics, typically to control for a variable or analyze patterns within specific segments of the data.

For instance, consider a scenario where you're examining the correlation between exercise frequency and weight loss across a diverse population. If you suspect that age could be a confounding variable impacting both exercise habits and weight loss outcomes, you might opt to stratify the data by age groups (e.g., 18-30, 31-50, 51-70). By doing so, you can then analyze the relationship between exercise frequency and weight loss within each age group separately, allowing for a more accurate assessment of the association between these variables while accounting for the potential influence of age.

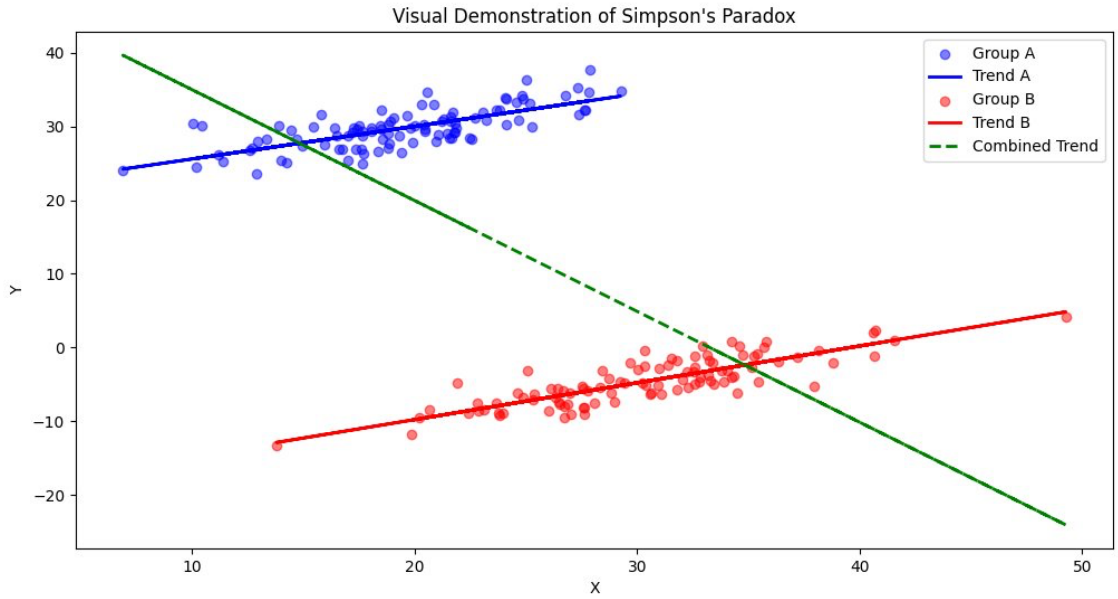
2.4 Simpson's Paradox in the Context of Numerical Variable

Expanding on the concept of Simpson's Paradox, it's important to recognize that its principle extends beyond categorical variables and also applies to numeric variables. When assessing the linear association between two numeric variables X and Y in isolation, the resulting conclusion may not only be directionally incorrect but could significantly distort the true effect of the factor under investigation.

For instance, consider a scenario where you're examining the relationship between temperature (X) and ice cream sales (Y). If you analyze their linear association without considering additional factors, you might observe a negative correlation, suggesting that as temperature increases, ice cream sales decrease. However, this conclusion might be misleading if you fail to account for confounding variables such as seasonality or time of day.

In such cases, overlooking important factors can obscure the true nature of the relationship between variables, leading to erroneous conclusions. Therefore, it's essential to conduct thorough analyses that consider potential confounders and contextual factors to ensure accurate interpretations of data trends and relationships.

Here's a visual representation illustrating the potential occurrence of Simpson's Paradox with numerical variables (source: https://twitter.com/PubliusVP/status/1758853645104214264/photo/1). Initially, when considering the entire dataset, variables X and Y might exhibit a negative correlation. However, upon stratifying the data into two distinct groups, A and B, we observe that within each subgroup, Y shows a positive correlation with X. This demonstration not only highlights the possibility of Simpson's Paradox affecting numerical variables but also underscores the effectiveness of stratification as a strategy for addressing it.



2.5 Key Takeaways

When examining the relationship between variables X and Y, it's imperative not to rely solely on their individual analysis. Doing so can lead to misleading conclusions, possibly even in the wrong direction. In such cases, attempting to quantify the extent of X's impact on Y becomes futile.

Instead, when assessing the influence of X on Y, it's essential to explore other potential factors that could affect Y. This is especially crucial if X isn't the primary driver of Y. By considering a broader range of variables, we gain a more comprehensive understanding of the complex interactions at play and can draw more accurate conclusions about the true relationship between X and Y.