Exercise Walkthrough: Spurious Correlation and Confounding

Justin Lanfermann

25. June 2025

Abstract

This document provides a detailed, step-by-step walkthrough for an exercise on correlation, covariance, and confounding variables. The solution is based on the concepts and notation introduced in the "Discrete Probability Theory" script by Niki Kilbertus (Summersemester 2025). Each step is explained with reference to the relevant definitions and theorems from the script to ensure a clear and rigorous understanding.

1 The Exercise

A data scientist is investigating the relationship between daily ice cream sales (I) and the number of drowning incidents (D) in a coastal town during the summer. They are given the following joint probability mass function (pmf) based on observed data, using simplified categories (Low/High).

	I = Low	**I = High**	**P(D = d)**
D = Low	0.35	0.10	0.45
D = High	0.05	0.45	0.55
P(I = i)	0.40	0.60	1.00

To simplify calculations, we assign numerical values: Low=0, High=1 for both I and D.

(i) Calculate $\mathbb{E}[I]$, $var[I]$, $\mathbb{E}[D]$, $var[D]$, $cov[I, D]$, and $\rho[I, D]$.

(ii) Based on $\rho[I, D]$, what do you conclude about the relationship between ice cream sales and drowning incidents?

(iii) A meteorologist suggests that the average daily temperature (T) is a confounding variable. Explain conceptually how a third variable could "create" the observed correlation between I and D. How would you check this, and what data would you need? How does this link to the famous saying that "correlation does not imply causation?"

2 Part (i): Calculating the Descriptive Statistics

2.1 Step 1: Understanding the Model and Data

We are given two random variables, I for ice cream sales and D for drowning incidents. Since they can only take values 0 (Low) and 1 (High), they are **Bernoulli random variables** [1]. The table provides their **joint probability mass function (pmf)**[3], $p_{I,D}(i, d) = P(I = i, D = d)$. The margins of the table, which we've filled in, provide the **marginal pmfs**[3], $p_I(i)$ and $p_D(d)$.

2.2 Step 2: Calculating Marginals and Expectations

The marginal pmfs are required to compute the expectation and variance of each variable individually. They are found by summing the joint probabilities over the other variable, as per *(Theorem 1.63, p. 35)*.

- $p_I(0) = P(I=0) = P(I=0, D=0) + P(I=0, D=1) = 0.35 + 0.05 = 0.40$
- $p_I(1) = P(I=1) = P(I=1, D=0) + P(I=1, D=1) = 0.10 + 0.45 = 0.60$
- $p_D(0) = P(D=0) = P(I=0, D=0) + P(I=1, D=0) = 0.35 + 0.10 = 0.45$
- $p_D(1) = P(D=1) = P(I=0, D=1) + P(I=1, D=1) = 0.05 + 0.45 = 0.55$

These match the totals provided in the table. Now we can compute the expectations using *(Definition 2.1, p. 42)*. For a discrete variable X, $\mathbb{E}[X] = \sum x \cdot p_X(x)$.

Expectation of I:

$$\mathbb{E}[I] = (0 \cdot p_I(0)) + (1 \cdot p_I(1)) = (0 \cdot 0.40) + (1 \cdot 0.60) = \mathbf{0.60}$$

Expectation of D:

$$\mathbb{E}[D] = (0 \cdot p_D(0)) + (1 \cdot p_D(1)) = (0 \cdot 0.45) + (1 \cdot 0.55) = \mathbf{0.55}$$

2.3 Step 3: Calculating Variances

We use the computational formula for variance from *(Remark 2.6, p. 45)*: $\text{var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. For a Bernoulli random variable that only takes values 0 or 1, we have $X^2 = X$, which means $\mathbb{E}[X^2] = \mathbb{E}[X]$. This simplifies the calculation significantly.

Variance of I:

$$\mathbb{E}[I^2] = \mathbb{E}[I] = 0.60$$
$$\text{var}[I] = \mathbb{E}[I^2] - (\mathbb{E}[I])^2 = 0.60 - (0.60)^2 = 0.60 - 0.36 = \mathbf{0.24}$$

Variance of D:

$$\mathbb{E}[D^2] = \mathbb{E}[D] = 0.55$$
$$\text{var}[D] = \mathbb{E}[D^2] - (\mathbb{E}[D])^2 = 0.55 - (0.55)^2 = 0.55 - 0.3025 = \mathbf{0.2475}$$

2.4 Step 4: Calculating Covariance and Correlation

To calculate the covariance, we first need $\mathbb{E}[ID]$. This expectation is calculated over the joint distribution, as per the Law of the Unconscious Statistician *(Lemma 2.2, p. 43)*.

$$\mathbb{E}[g(I, D)] = \sum_{i,d} g(i, d) \cdot p_{I,D}(i, d)$$

Here, $g(I, D) = ID$. The product ID is only non-zero when both $I = 1$ and $D = 1$.

$$\mathbb{E}[ID] = (0 \cdot 0 \cdot 0.35) + (0 \cdot 1 \cdot 0.05) + (1 \cdot 0 \cdot 0.10) + (1 \cdot 1 \cdot 0.45) = 0.45$$

Now we use the computational formula for covariance from *(Remark 2.10, p. 46)*: $\text{cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

$$\text{cov}[I, D] = \mathbb{E}[ID] - \mathbb{E}[I]\mathbb{E}[D] = 0.45 - (0.60)(0.55) = 0.45 - 0.33 = \mathbf{0.12}$$

Finally, we calculate the Pearson correlation coefficient using *(Definition 2.13, p. 48)*:

$$\rho[I, D] = \frac{\text{cov}[I, D]}{\sqrt{\text{var}[I]\text{var}[D]}} = \frac{0.12}{\sqrt{0.24 \cdot 0.2475}} = \frac{0.12}{\sqrt{0.0594}} \approx \frac{0.12}{0.2437} \approx \mathbf{0.4924}$$

3 Part (ii): Interpreting the Correlation

The correlation coefficient $\rho[I, D] \approx 0.4924$ indicates a **moderate positive linear relationship** between ice cream sales and drowning incidents.

Reasoning: According to *(Proposition 2.14, p. 48)*, the correlation coefficient ρ is always between -1 and 1.

- A value of $\rho > 0$ indicates a positive association: when one variable is high, the other tends to be high as well.
- A value close to 1 indicates a strong linear relationship, while a value close to 0 indicates a weak or no linear relationship.

Our value of ≈ 0.49 is positive and reasonably far from zero, suggesting that days with higher ice cream sales do indeed tend to be days with a higher number of drowning incidents. Looking at the joint pmf confirms this: the probability of both being high ($P(I = 1, D = 1) = 0.45$) is much larger than what would be expected if they were independent ($P(I = 1)P(D = 1) = 0.60 \cdot 0.55 = 0.33$).

4 Part (iii): The Specter of Confounding

4.1 The Concept of Confounding

The meteorologist's suggestion of temperature (T) as a confounding variable is the key to understanding the situation. A **confounding variable** is a third variable that influences both the "cause" (here, I) and the "effect" (here, D), creating a **spurious correlation**[2].

Conceptual Explanation: The observed correlation between I and D might not be because ice cream causes drowning. Instead, a third factor—high temperature—could be the **common cause**:

1. **High Temperature → High Ice Cream Sales:** On hot days, more people buy ice cream. (T influences I).
2. **High Temperature → High Drowning Incidents:** On hot days, more people go swimming, which naturally increases the total number of people in the water and thus the opportunity for drowning incidents. (T influences D).

Because both I and D are effects of the common cause T, they move together. They appear correlated, but this correlation is non-causal. The relationship is not $I \to D$, but rather $I \leftarrow T \to D$.

4.2 How to Check for Confounding

To verify this hypothesis, we need to check the relationship between I and D *after* accounting for, or "controlling for," the effect of T. The formal tool for this is the **partial correlation coefficient**, $\rho[I, D|T]$, as introduced in *(Definition 2.16, p. 49)*.

This coefficient measures the linear relationship between I and D while holding T constant. If temperature is indeed the sole confounder, we would expect to find:

$$\rho[I, D|T] \approx 0$$

This would mean that for a *given* temperature (e.g., looking only at days where it was 30°C), there is no significant relationship between ice cream sales and drownings.

Data Requirements: To calculate this, we would need a more detailed dataset that includes temperature. Specifically, we would need the **joint pmf** $p_{I,D,T}(i, d, t)$. This means we would need to stratify our original table by temperature levels (e.g., "Cool", "Warm", "Hot") and then apply the formula for partial correlation.

4.3 Link to "Correlation Does Not Imply Causation"

This example is the classic illustration of the maxim that **correlation does not imply causation**[4].

- We found a **correlation**: a real statistical association exists in the data between I and D.
- This does not mean we've found a **causal link**. It's highly unlikely that buying ice cream makes people more likely to drown, or that the grief from drownings drives ice cream sales.

The presence of a confounding variable (T) provides a plausible alternative explanation for the observed correlation. Without controlling for potential confounders, one can easily draw incorrect causal conclusions from purely observational data. Establishing causation is a much harder problem that requires either carefully designed experiments (like randomized controlled trials) or advanced causal inference methods.

5 Further Concepts

[1] Bernoulli Random Variable

A random variable X is called a Bernoulli random variable with parameter $p \in [0, 1]$ if it can only take two values, typically 0 ("failure") and 1 ("success"). Its probability mass function (pmf) is given by:

$$p_X(k) = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \end{cases}$$

This is a specific case of the Bernoulli process with $n = 1$ *(Example 1.36(iii), p. 18)*. Its key properties are:

- **Expectation:** $\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1 - p) = p$
- **Variance:** $\text{var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1 - p)$

In our exercise, $I \sim \text{Ber}(0.60)$ and $D \sim \text{Ber}(0.55)$.

[2] Spurious Correlation

A spurious correlation is a statistical relationship in which two or more variables are associated, but not causally related. The association arises either due to coincidence or, more commonly, due to the presence of a third, unobserved factor known as a **confounding variable** or a "common cause". The ice cream sales and drowning example is a textbook case where temperature is the confounder creating a spurious correlation.

[3] Joint and Marginal Distributions

For a set of random variables (e.g., X, Y), their **joint distribution** *(Def 1.61, p. 34)* describes the probability of them taking on a particular combination of values simultaneously, i.e., $P(X = x, Y = y)$.

The **marginal distribution** of one variable (e.g., X) is its probability distribution irrespective of the other variables. It is obtained by "summing out" (or integrating out for continuous variables) the other variables from the joint distribution, as described in *(Thm 1.63, p. 35)*.

$$P(X = x) = \sum_y P(X = x, Y = y)$$

[4] Correlation vs. Causation

This is a fundamental principle in statistics and scientific reasoning.

- **Correlation** is a statistical measure that indicates the extent to which two variables fluctuate together. It describes an association.
- **Causation** indicates that one event is the result of the occurrence of the other event; i.e., there is a mechanism by which one variable directly influences the other.

The existence of a correlation is a necessary condition for a direct causal relationship, but it is not a sufficient one. Spurious correlations due to confounding are one of the primary reasons why this is true. To establish causation, one typically needs controlled experiments or more sophisticated causal inference techniques, a major topic in machine learning and statistics.