Covariance and Correlation

Gregory M. Shinault

Goal

The joint PMF is complete information about the relationship between X and Y. Unfortunately it can be difficult to interpret the relationship between X and Y from the PMF.

Covariance and correlation provide a simple way to interpret this relationship.

This material corresponds to section 8.4 of the textbook.

Covariance

Definition

The *covariance* of random variables X and Y is given by

$$\operatorname{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

If $\operatorname{Cov}(X, Y) > 0$, we say X and Y are *positively correlated*.
If $\operatorname{Cov}(X, Y) < 0$, we say X and Y are *negatively correlated*.
If $\operatorname{Cov}(X, Y) = 0$, we say X and Y are *uncorrelated*.

Computation

Fact: The *covariance* of random variables X and Y is computed by

$$\operatorname{Cov}(X, Y) = \mathbb{E}XY - \mu_X \mu_Y.$$

Example

Suppose X and Y have the joint PMF

		Y		
		1	2	3
X	1	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{4}$
	2	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$

Find $\operatorname{Cov}(X, Y)$.

Example

Suppose (X, Y) are uniformly distributed on the circle of radius 2 at the origin. Find $\text{Cov}(X, Y)$.

Properties

1. Symmetry: $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.
2. Bilinearity:

$$\begin{aligned}&\text{Cov}(aX + bY, cW + dZ)\\&= ac\text{Cov}(X, W) + ad\text{Cov}(X, Z) + bc\text{Cov}(Y, W) + bd\text{Cov}(Y, Z).\end{aligned}$$

Properties

Fact: The variance of a sum of random variables can be found with the formula

$$\text{Var}(X_1 + X_2 + \cdots + X_n) = \sum_{j=1}^{n} \text{Var}(X_j) + 2\sum_{i<k} \text{Cov}(X_i, X_k).$$

Special Example | Hypergeometric Distribution

Suppose $X \sim \text{HyperGeo}(N, N_A, n)$. Find the variance of X.

Shortcomings

1. The magnitude of the covariance is not indicative of the strength of the relationship between X and Y. (Changing units changes the covariance, but the underlying relationship should not change)
2. Covariance only measures the linear relationship between X and Y. (We will not address this issue)

Correlation

Definition

Let

$$X_* = \frac{X - \mu_X}{\sigma_X}, \quad Y_* = \frac{Y - \mu_Y}{\sigma_Y}.$$

The *correlation* of X and Y is defined as

$$\text{Corr}(X, Y) = \text{Cov}(X_*, Y_*) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Key Properties

1. $-1 \leq \text{Cor}(X, Y) \leq 1$.
2. $\text{Corr}(X, Y) = 1$ if and only if $Y = aX + b$ for some positive a.
3. $\text{Corr}(X, Y) = -1$ if and only if $Y = -aX + b$ for some positive a.

Special Example | Multinomial Distribution

Suppose $(X_1, \ldots, X_r) \sim \text{Mult}(n, p_1, \ldots, p_r)$. Find $\text{Corr}(X_i, X_j)$.

Summary

Key Ideas

1. $\text{Cov}(X, Y) = \mathbb{E}XY - \mu_X \mu_Y$.
2. $\text{Cor}(X, Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$.
3. Both are used to measure the linear relationship between X and Y.
4. They possess many properties. You must know them all.