Properties of the Correlation Coefficient via a Scalar Product in the Vector Space of Random Variables

Horst Hohberger

JOINT INSTITUTE
交大密西根学院

1 Vector Space Structure

Given a sample space $(S, \mathcal{F}, P)$ we may consider the set of all random variables

$$X\colon S \to \mathbb{R}$$

on that space. Let us denote this set by V. Then V contains, for example, the *constant random variables*

$$X_c\colon S \to \mathbb{R}, \qquad X_c(p) = c \quad \text{for all } p \in S,$$

where $c \in \mathbb{R}$ is a fixed constant (any element of the sample space is mapped to the same number c.). For simplicity, we often denote X_c simply by c. We can add two elements $X, Y \in V$ by defining

$$(X+Y)(p) := X(p) + Y(p)$$

and we can multiply with a real number $\lambda \in \mathbb{R}$ by setting

$$(\lambda X)(p) = \lambda \cdot X(p).$$

(These are the usual *point-wise* definitions for sums and scalar multiples of functions.) In this way, V becomes a real vector space.

If we want to have V include only discrete or only continuous random variables, we can modify these definitions accordingly.

Remark 1. We can also multiply two random variables, defining

$$(X \cdot Y)(p) := X(p) \cdot Y(p).$$

2 Expectation

For a discrete random variable X with density f_X we define the *expectation*

$$\mathrm{E}[X] := \sum_{x \in \operatorname{ran} X} x \cdot f_X(x) \tag{1}$$

and, more generally, if $g\colon \mathbb{R} \to \mathbb{R}$ is a continuous function, we find that

$$\mathrm{E}[g(X)] = \sum_{x \in \operatorname{ran} X} g(x) \cdot f_X(x). \tag{2}$$

If X is a continuous random variable, the analogous definiton is

$$\mathrm{E}[g(X)] := \int_{\mathbb{R}} g(x) \cdot f_X(x)\, dx.$$

Remark 2. Of course, $Y = g(X)$ is a random variable in its own right with density f_Y. It can be shown that $\mathrm{E}[g(x)]$ by (2) and $\mathrm{E}[Y]$ by (1) coincide.

We immediately see that if X is a constant random variable,

$$\mathrm{E}[X_c] = c$$

and that

$$\mathrm{E}[\lambda \cdot X] = \lambda \cdot \mathrm{E}[X].$$

Given two random variables X and Y (either both discrete or both continuous), it can be shown that

$$\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y].$$

We postpone the proof of this relation.

3 Almost Sure Equality and Partition of V

We will say that two random variables X and Y are equal *almost surely* if

$$P[X \le x] = P[Y \le x] \qquad \text{for all } x \in \mathbb{R}.$$

We then write $X = Y$ a.s.
In particular, if X and Y are discrete random variables with common domain $\Omega \subset \mathbb{R}$ and densities f_X and f_Y, respectively, this implies

$$\sum_{z \le x} f_X(z) = \sum_{z \le x} f_Y(z).$$

from which we can deduce $f_X(x) = f_Y(x)$ for all $x \in \Omega$.
If X and Y are continuous random variables, then $X = Y$ almost surely means that

$$\int_{-\infty}^{x} f_X(z)\,dz = \int_{-\infty}^{x} f_Y(z)\,dz$$

in which case f_X and f_Y may differ on sets of measure zero.
We can define an equivalence relation $\sim$ on V by saying that X and Y are equivalent ($X \sim Y$) if they are equal almost surely. We then denote the set of equivalence classes (partition) by $V/\sim$, which is of course again a vector space.

4 Chebyshev Inequality

A random variable X is said to be *positive* if $P[X < 0] = 0$. In that case, $f_X(x) = 0$ for $x < 0$.
Example 1. If X is a random variable, then $Y = X^2$ is positive, since

$$P[Y \le y] = P[X^2 \le y] = \begin{cases} 0 & y < 0 \\ P[-\sqrt{y} \le X \le \sqrt{y}] & \text{otherwise.} \end{cases}$$

Let X be a positive random variable and $y > 0$ be fixed. Then

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_{0}^{\infty} x f_X(x)\,dx \\ &\ge \int_{y}^{\infty} x f_X(x)\,dx \ge \int_{y}^{\infty} y f_X(x)\,dx \\ &= y \cdot P[X \ge y] \end{aligned}$$

Hence, we obtain a version of the Chebyshev inequality,

$$P[X \ge y] \le \frac{E[X]}{y}$$

We would like to prove the following result:
Lemma 1. *Let X be a positive random variable and suppose that $E[X] = 0$. Then $X = 0$ almost surely.*

Proof. We first note that

$$P[X < 0] + P[X = 0] + P[X > 0] = 1.$$

Since $P[X < 0] = 0$ by assumption, it remains to show $P[X > 0] = 0$. Then $P[X = 0] = 1$, which means that $X = 0$ almost surely.
Note that $X > 0$ is equivalent to stating that $X > 1/n$ for some $n \in \mathbb{N} \setminus \{0\}$ and so

$$\{p \in S \colon X(p) > 0\} = \bigcup_{n \in \mathbb{N}\setminus\{0\}} \{p \in S \colon X(p) > 1/n\}$$

By Boole's inequality ($P[A \cup B] \le P[A] + P[B]$), this implies that

$$P[X > 0] \le \sum_{n=1}^{\infty} P[X > 1/n].$$

Applying Chebyshev's inequality,

$$P[X \ge 1/n] \le \frac{E[X]}{y} = 0.$$

and so $P[X > 0] = 0$. □

5 A Scalar Product on $V/\sim$

Recall that a scalar product on a real vector space V is a map $\langle\,\cdot\,,\,\cdot\,\rangle\colon V \to \mathbb{R}$ with the properties

i) $\langle u, u\rangle \geq 0$,

ii) $\langle u, u\rangle = 0$ if and only if $u = 0$,

iii) $\langle u, v\rangle = \langle v, u\rangle$,

iv) $\langle u, \lambda v\rangle = \lambda \cdot \langle u, v\rangle$,

v) $\langle u, v + w\rangle = \langle u, v\rangle + \langle u, w\rangle$

for all $u, v, w \in V$ and $\lambda \in \mathbb{R}$.
We now claim that

$$\langle X, Y\rangle := \mathrm{E}[XY] \tag{3}$$

defines a scalar product on the set $V/\sim$. We will show this for the case of V being the set of continuous random variables only; the discrete case is completely analogous.

i) $\langle X, X\rangle := \mathrm{E}[X^2] = \int_{\mathbb{R}} x^2 f_X(x)\,dx \geq 0$ since $f_X(x) \geq 0$ for all $x \in \mathbb{R}$.

ii) $X = 0$ implies $\mathrm{E}[X^2] = \mathrm{E}[0] = 0$. Conversely, $\mathrm{E}[X^2] = 0$ implies $X = 0$ almost surely (since X^2 is positive, $X^2 = 0$ almost surely by Lemma 1 and hence $X = 0$ almost surely). But then $X = 0$ in $V/\sim$.

iii) $\mathrm{E}[XY] = \mathrm{E}[YX]$

iv) $\mathrm{E}[X(\lambda Y)] = \mathrm{E}[\lambda XY] = \lambda\,\mathrm{E}[XY]$

v) $\mathrm{E}[X(Y + Z)] = \mathrm{E}[XY + XZ] = \mathrm{E}[XY] + \mathrm{E}[XZ]$.

Since (3) defines a scalar product on a vector space, we know that the Cauchy-Schwartz inequality holds,

$$\mathrm{E}[XY]^2 \leq \mathrm{E}[X^2] \cdot \mathrm{E}[Y^2]$$

and, in particular,

$$\mathrm{E}[XY]^2 = \mathrm{E}[X^2] \cdot \mathrm{E}[Y^2] \qquad \text{if and only if} \qquad X = \lambda \cdot Y \quad \text{a.s. for some } \lambda \in \mathbb{R}.$$

Remark 3. Should we need to recall a proof of the Cauchy-Schwartz inequality, here is one:
Let u, v be arbitrary vectors in a vector space. Write $\|u\| := \sqrt{\langle u, u\rangle}$, $e := v/\|v\|$. Then $\langle e, e\rangle = \langle v, v\rangle/\|v\|^2 = 1$ and

$$\begin{aligned} 0 \leq \|u - \langle e, u\rangle e\|^2 &= \langle u - \langle e, u\rangle e, u - \langle e, u\rangle e\rangle \\ &= \|u\|^2 - |\langle e, u\rangle|^2 \end{aligned}$$

It follows that

$$|\langle u, v\rangle|^2 = \|v\|^2 \cdot |\langle u, e\rangle|^2 \leq \|u\|^2 \cdot \|v\|^2.$$

Suppose that

$$|\langle u, v\rangle|^2 = \|u\|^2 \cdot \|v\|^2.$$

This is equivalent to $|\langle e, u\rangle|^2 = \|u\|^2$ or $\|u - \langle e, u\rangle e\| = 0$, which in turn means that

$$u = \langle e, u\rangle e = \frac{\langle v, u\rangle}{\langle v, v\rangle} v.$$

6 Variance and Covariance

From the point of view of linear algebra, $\mathrm{E}[X^2]$ defines the square of the norm of X, i.e., it gives a measure of the size of X. However, this has little meaning in probability theory. What is of interest, instead, is the size of the difference between X and its mean $\mu := \mathrm{E}[X]$. Therefore, we define

$$\mathrm{Var}[X] := \mathrm{E}[(X - \mu)^2]$$

(which is the square of the norm of the difference between X and the constant random variable X_μ.) The standard deviation σ is defined to be the positive square root of the variance and is therefor analogous to the norm of $X - \mu$. We often use σ^2 to denote the variance.

Remark 4. In probability theory, taking an analogy to mechanics, one calls $\mathrm{E}[X]$ the *first moment* of X, $\mathrm{E}[X^2]$ the *second moment* and generally $\mathrm{E}[X^k]$ the kth moment of X. When the mean is subtracted, the corresponding moment is said to be *centered*. Trivially, the centered first moment vanishes, $\mathrm{E}[X - \mu] = 0$. Hence, the variance is the centered second moment.

It is sometimes useful to note that

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2.$$

and that

$$\mathrm{Var}[\lambda \cdot X] = \lambda^2 \cdot \mathrm{Var}[X], \qquad \mathrm{Var}[X_c] = 0 \qquad \text{for a constant random variable } X_c.$$

Given two random variables X and Y, we define the covariance of X and Y through the scalar product

$$\langle X - \mu_X, Y - \mu_Y \rangle = \mathrm{E}\big[(X - \mu_X)(Y - \mu_Y)\big] =: \mathrm{Cov}[X, Y].$$

It is then easy to check that

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Cov}[X, Y].$$

as would be expected from linear algebra. As for the variance, it is not difficult to show that

$$\mathrm{Cov}[X, Y] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y].$$

We standardize a random variable by subracting its mean and dividing by the standard deviation, where we use that

$$\mathrm{Var}[X - \mu] = \mathrm{Var}[X] = \sigma^2,$$

i.e., a standardized X is given by

$$\frac{X - \mu}{\sigma}.$$

This corresponds to a normalized vector (of unit lenth) in linear algebra. Note that $\frac{X-\mu}{\sigma}$ has mean zero and unit variance, so

$$\mathrm{E}\left[\left(\frac{X - \mu}{\sigma}\right)^2\right] = \mathrm{Var}\left[\frac{X - \mu}{\sigma}\right] + \mathrm{E}\left[\frac{X - \mu}{\sigma}\right]^2 = 1.$$

7 Correlation Coefficient

In linear algebra, the scalar product is a measure for the angle between vectors. In probability theory, the corresponding concept is called *correlation*. Given two random variables X and Y, we define the correlation coefficient ϱ_{XY} by

$$\varrho_{XY} := \mathrm{Cov}\left[\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right] = \frac{\mathrm{Cov}[X, Y]}{\sigma_X \sigma_Y}$$

From the Cauchy-Schwartz inequality, we see immediately that

$$\varrho_{XY}^2 \le \mathrm{E}\left[\left(\frac{X - \mu_X}{\sigma_X}\right)^2\right] \mathrm{E}\left[\left(\frac{Y - \mu_Y}{\sigma_Y}\right)^2\right] = 1$$

and $\varrho_{XY}^2 = 1$ if and only if

$$\frac{X - \mu_X}{\sigma_X} = \lambda \cdot \frac{Y - \mu_Y}{\sigma_Y} \qquad \text{a.s. for some } \lambda \in \mathbb{R}.$$

Hence, $|\varrho_{XY}| = 1$ of and only if for some $\beta_0, \beta_1 \in \mathbb{R}$ we have

$$Y = \beta_1 X + \beta_0$$

almost surely.