Contents

(All variables are real and one-dimensional unless otherwise specified.)

1. Mean, variance, and moment

The **mean** of a random variable X is

$$\langle X\rangle = \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x\,,$$

where $f_X(x)$ is its probability density function (PDF). A valid PDF must satisfy $\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$ so that we can always find X somewhere on the number line. Its **variance** takes a similar form:

$$\begin{aligned}\sigma_X^2 &= \langle (X-\langle X\rangle)^2\rangle \\ &= \langle X^2\rangle - \langle X\rangle^2\,.\end{aligned}$$

In general, $\langle X^n\rangle$ is called the nth **moment** of X, whereas $\langle (X-\langle X\rangle)^n\rangle$ is called its nth **central moment**. The third and the fourth moments of a random variable are related to its **skewness** and **kurtosis** respectively.

2. Covariance and correlation

The **covariance** between two random variables X and Y reads

$$\begin{aligned}\sigma_{XY}^2 &= \langle (X - \langle X \rangle)\,(Y - \langle Y \rangle) \rangle \\ &= \langle XY \rangle - \langle X \rangle \langle Y \rangle\,.\end{aligned}$$

It gets this name because it is formally identical to variance. (If $X = Y$, $\sigma_{XY}^2 = \sigma_X^2$.) Two variables has a positive covariance if an increase in one often occurs with an increase in the other. On the other hand, two variables has a more negative covariance if they possess opposite trends.

One would usually like to normalize covariance as **correlation** (or correlation coefficient) to compare behaviours of various pairs of variables. The correlation between X and Y is canonically defined as

$$r_{XY} = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y} \in [-1, 1]\,,$$

which may be specifically called **Pearson's correlation**. (I will soon introduce two more kinds of correlation.)

2.1 Correlation versus dependence

The correlation between two variables **only** indicates the **strength** of their **linear** dependence. In other words, a higher correlation between two variable means their scatter plot resemble a straight line more **regardless of its slope.** Fig. 1 well illustrates this point.

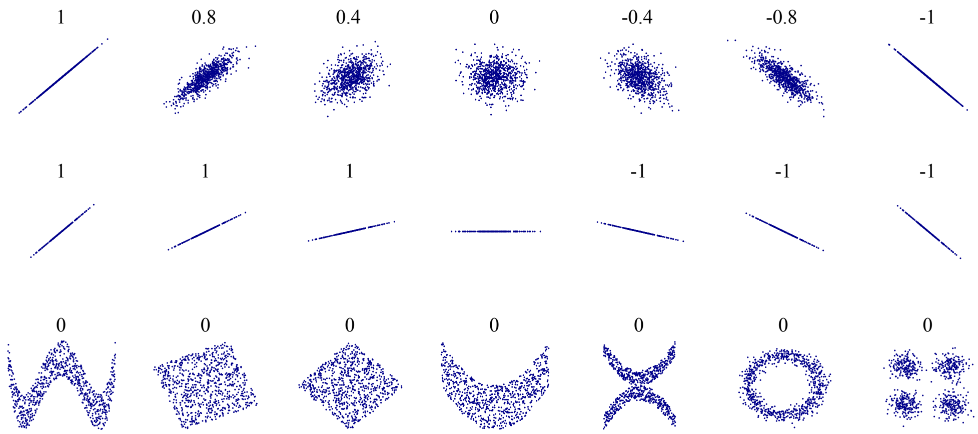

Fig. 1 The number above each subplot indicates the correlation between the horizontal and the vertical variables. (Retrieved from Wikimedia Commons.) For the central subplot, correlation is undefined for the zero variance of its vertical variable.

2.2 Example: quadratic dependence

Let $X \sim \mathcal{U}(-1,1)$ be a random variable, i.e. X is uniformly distributed in $[-1,1]$. What is its correlation between X and $Y=X^2$?

Solution. Because of its uniform distribution, the PDF of X is $f_X(x)=\begin{cases}\frac{1}{2} & (-1\le x\le 1)\\ 0 & (\text{otherwise})\end{cases}$. Then we can compute $\langle X\rangle$ and thus $\sigma^2_{XY}=\langle XY\rangle-\langle X\rangle\langle Y\rangle=\langle X^3\rangle-\langle X\rangle\langle X^2\rangle$.

$$\begin{aligned}\langle X\rangle &= \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x\\ &= \frac{1}{2}\int_{-1}^{1} x\mathrm{d}x = 0\end{aligned}$$

Similarly, $\langle X^3 \rangle = 0$, so σ^2_{XY} also vanishes. This result implies that the correlation between X and Y is $r_{XY} = 0$, and it ultimately teaches us that (Pearson's) correlation does not measure the strength of nonlinear dependence reliably.

3. Rank correlation

Two useful alternatives to Pearson's correlation are **Spearman's correlation** and **Kendall's correlation**. They can be collectively called **rank correlation** and are devised to respond more sensitively to nonlinear dependence.

Unlike Pearson's correlation, rank correlation usually requires an **explicit** knowledge of observed data, so let us first assume there are n realizations of (X, Y), i.e. $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. Rank correlation first transforms each realized pair (x_i, y_i) to a pair of **rank variables** (R_{x_i}, R_{y_i}), where $R_{x_i} = k$ if x_i is the kth smallest realization of X.

Spearman's correlation. Also called Spearman's ρ. It is in fact the Pearson's correlation between $R_X = \{R_{x_i}\}$ and $R_Y = \{R_{y_i}\}$, so

$$\rho_{XY} = \frac{\sigma^2_{R_X R_Y}}{\sigma_{R_X} \sigma_{R_Y}} .$$

Kendall's correlation. Also called Kendall's τ. It first assigns two scores $\hat{x}_{ij}$ and $\hat{y}_{ij}$ to each pair of (R_{x_i}, R_{y_i}) and $\left(R_{x_j}, R_{y_j}\right)$ having $j > i$.

$$\begin{cases} \hat{x}_{ij} = \operatorname{sgn}\left(R_{x_i} - R_{x_j}\right) \\ \hat{y}_{ij} = \operatorname{sgn}\left(R_{y_i} - R_{y_j}\right) \end{cases}$$

Then the correlation is defined as

$$\tau_{XY} = \frac{2}{n\left(n-1\right)} \sum_{i<j} \hat{x}_{ij} \hat{y}_{ij} \,.$$

In Kendall's original terms, a pair is **concordant** if $\hat{x}_{ij}\hat{y}_{ij} > 0$ but **discordant** if $\hat{x}_{ij}\hat{y}_{ij} < 0$, and it is neither concordant nor discordant if $\hat{x}_{ij}\hat{y}_{ij} = 0$. Fig. 2 illustrates this idea.

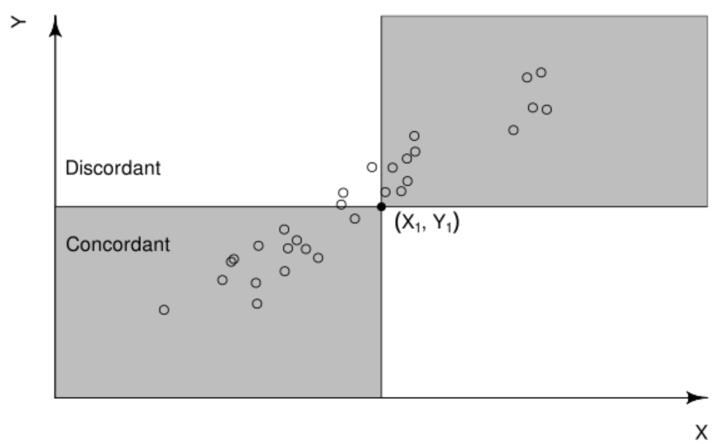

Fig. 2 The point (X_1, Y_1) forms concordant pairs with the points in the grey regions but discordant pairs with the ones in the white regions. (Retrieved from Wikimedia Commons.)

The concepts of concordance and discordance help reformulate Kendall's correlation as

$$\tau_{XY} = \frac{\#\text{concordant pairs} - \#\text{discordant pairs}}{n\left(n-1\right)/2}.$$

4.1 Example: revisiting quadratic dependence

Let $X \sim \mathcal{U}\left(0,1\right)$ be a random variable, i.e. X is uniformly distributed in $\left[0,1\right]$. What are its Pearson's correlation and Spearman's correlation with $Y = X^2$?

Solution. The Pearson's correlation is $r_{XY} = \frac{\langle X^3\rangle - \langle X\rangle\langle X^2\rangle}{\sqrt{\langle X^2\rangle - \langle X\rangle^2}\sqrt{\langle X^4\rangle - \langle X^2\rangle^2}}$, which can be shown to equal $\frac{\sqrt{15}}{4} \approx 0.968$. The Spearman's correlation is simply $\rho_{XY} = 1$ because a larger realization of $X \in \left[0,1\right]$ definitely leads to a larger Y.