Stat 400, section 5.2 Expected Values, Covariance and Correlation

notes by Tim Pilachowski

Example A: In the game of Chinchillas and Caves™ (C&C™) there are two four-sided dice (triangular pyramids) which determine a player's next move. The purple die (X) has one side that has the number 0, one side with the number 1, and two sides with the number 2. The yellow die (Y) has sides numbered 0, 1, 1 and 3.

x \ y	0	1	3	$p_X(x) =$
0	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{4}$
1	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{4}$
2	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{2}$
$p_Y(y) =$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	

In odd-numbered months, the length of a player's move is determined by the sum of the numbers on the two dice (random variable $B = X + Y$). In even-numbered months on even-numbered dates, the length of a player's move is determined by the product of the numbers on the two dice (random variable $C = XY$). In even-numbered months on odd-numbered dates (with the exception of February 29 in a leap year), the length of a player's move is determined by the maximum of the two numbers shown on the dice (random variable $D = \max [X, Y]$). On February 29 in a leap year, players get to choose sum, product or maximum. Which choice would result in the largest average move?

$E(B) = ?$ method: Develop the probability distribution table for random variable B first, then calculate.

B	0	1	2	3	4	5
$p_B(b) =$						

Since 1) addition is commutative and associative, and since 2) multiplication is distributive over addition, when we found $E(B)$, we calculated $E(B) = \sum_{\text{all } b} b * p(b) = \sum_{\text{all } x} \sum_{\text{all } y} (x + y) * p(x, y) = \sum_{\text{all } x} \sum_{\text{all } y} b(x, y) * p(x, y)$.

$E(C) = ?$ method: Calculate [values of each of the nine possible products] times [nine associated probabilities].

$E(D) = ?$

The concept underlying calculation of expected value above can be carried over from discrete to continuous random variables.

discrete $$E[h(X,Y)] = \sum_{\text{all } x} \sum_{\text{all } y} h(x,\ y) * p(x,\ y)$$

continuous $$E[h(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,\ y) * f(x,\ y)\, dx\, dy$$

(See Lecture 3.6b in which the concept of transformation of a single random variable X was explored.)

Example B: Two continuous random variables X and Y have joint probability density function

$$f(x,\ y) = \begin{cases} \dfrac{1}{2xy} & 1 \le x \le e, 1 \le y \le e^2 \\ 0 & \text{otherwise} \end{cases}.$$

a) Given random variable $B = X + Y$, find $E(B)$.

b) Given random variable $C = XY$, find $E(C)$.

The variance of a single random variable X gives an indication of how the values vary in relationship to the mean. Given two random variables X and Y, we'll be interested in how the two vary in relationship to each other. The *covariance* between two random variables is

discrete $$\text{Cov}(X,Y) = \sum_{\text{all } x} \sum_{\text{all } y} (x - \mu_X) * (y - \mu_Y) * p(x,\ y)$$

continuous $$\text{Cov}(X,Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X) * (y - \mu_Y) * f(x,\ y)\, dx\, dy$$

However, as with single random variables, these calculations can become quite onerous. If we were to multiply $(x - \mu_X) * (y - \mu_X)$, then find expected value for each term separately before recombining, we'd get a shortcut:

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

Note:

Example A revisited: In the game of Chinchillas and Caves™ (C&C™) there are two four-sided dice (triangular pyramids) which determine a player's next move. The purple die (X) has one side that has the number 0, one side with the number 1, and two sides with the number 2. The yellow die (Y) has sides numbered 0, 1, 1 and 3.

x \ y	0	1	3	$p_X(x) =$
0	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{4}$
1	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{4}$
2	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{2}$
$p_Y(y) =$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	

Find Cov(X, Y).

Example B revisited: Two continuous random variables X and Y have joint probability density function $f(x, y) = \begin{cases} \frac{1}{2xy} & 1 \le x \le e, 1 \le y \le e^2 \\ 0 & \text{otherwise} \end{cases}$. Find Cov($X$, Y).

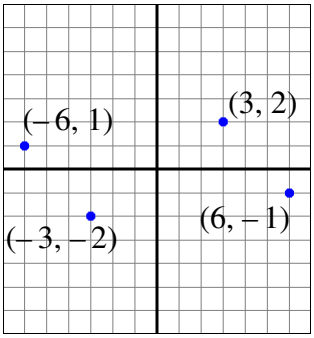
Example C: Theorem: If two random variables X and Y are independent, then Cov(X, Y) = 0.

Proof:

Note that this is a one-way conditional statement. "Cov(X, Y) = 0" does not imply independence.

Example D: Two discrete random variables have joint probability distribution

$$p(x, y) = \begin{cases} \frac{1}{4} & (x, y) = (-6, 1), (-3, -2), (3, 2), (6, -1) \\ 0 & \text{otherwise} \end{cases}.$$



	− 2	− 1	1	2	$p_X(x) =$
− 6					
− 3					
3					
6					
$p_Y(y) =$					

Are X and Y independent?

$E(X) =$

$E(Y) =$

$E(XY) =$

$\text{Cov}(X, Y) =$

Example E: Two continuous random variables X and Y have joint probability density function

$$f(x, y) = \begin{cases} \frac{x + y}{3} & 0 \le x \le 1, 0 \le y \le 2 \\ 0 & \text{otherwise} \end{cases}.$$ Find $\text{Cov}(X, Y)$.

Covariance of two random variables has a basic problem that it shares with variance of one random variable: the value alone doesn't tell us much. Given a random variable X with variance $\sigma_X^2 = 10$, and a linear transformation $Y = 5X$, then $\sigma_Y^2 = 5^2 * \sigma_X^2 = 250$. In other words, the size of the values of X has a direct effect on the values of $E(X)$ and $V(X)$.

For one random variable, we found standard deviation, then used $Z = \dfrac{X - \mu_X}{\sigma_X}$ to standardize. We'll do something similar for two random variables considered jointly. The *correlation coefficient* of two random variables X and Y is defined as $\text{Corr}(X, Y) = \rho_{X,Y} = \rho = \dfrac{\text{Cov}(X, Y)}{\sigma_X * \sigma_Y}$.

Example A again: In the game of Chinchillas and Caves™ (C&C™) there are two four-sided dice (triangular pyramids) which determine a player's next move. The purple die (X) has one side that has the number 0, one side with the number 1, and two sides with the number 2. The yellow die (Y) has sides numbered 0, 1, 1 and 3. Find the correlation coefficient $\rho_{X,Y}$.

Example B again: Two continuous random variables X and Y have joint probability density function

$$f(x, y) = \begin{cases} \dfrac{1}{2xy} & 1 \le x \le e, 1 \le y \le e^2 \\ 0 & \text{otherwise} \end{cases}$$

. Find Corr(X, Y) = $\rho_{X,Y}$.

Example D again: Two discrete random variables have joint probability distribution

$$p(x, y) = \begin{cases} \dfrac{1}{4} & (x, y) = (-6, 1), (-3, -2), (3, 2), (6, -1) \\ 0 & \text{otherwise} \end{cases}$$

. Find ρ.

Note that in the case of Examples A and B, because the two variables are independent we could have known already that $\rho = 0$, since independence of X and Y implies Cov(X, Y) = 0 (see Example D) thus independence implies $\rho = 0$. However, the converse is not true, as in Example D: While Corr(X, Y) = ρ = Cov(X, Y) = 0, the variables X and Y are not independent.

Example E again: Two continuous random variables X and Y have joint probability density function

$$f(x, y) = \begin{cases} \dfrac{x + y}{3} & 0 \le x \le 1, 0 \le y \le 2 \\ 0 & \text{otherwise} \end{cases}$$

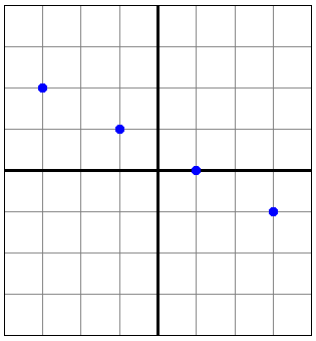
. Find Corr(X, Y) = $\rho_{X,Y}$.

Correlation tells us about the relationship or connection between X and Y. A positive value for ρ indicates that X and Y increase or decrease together. A negative value for ρ indicates that when X increases, Y decreases, and vice versa.

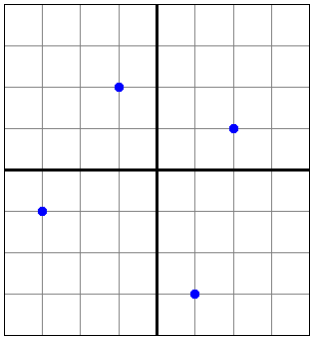
Height and weight have a positive correlation. Unemployment and inflation have a negative correlation.

The correlation coefficient is a measure of the *linear* relationship between random variables X and Y. (In an upper level stats course, you might investigate quadratic, cubic, logarithmic or exponential relationships.)

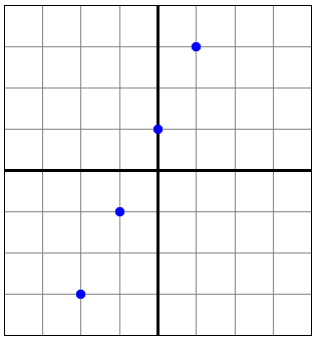
Proposition: $-1 < \rho < 1$.



$\rho = -1$
$Y = aX + b, A < 0$



$\rho = 0$



$\rho = 1$
$Y = aX + b, A > 0$

The sign of ρ indicates direction of the correlation; the magnitude of ρ indicates strength of the correlation. This text's rule of thumb: $|\rho| < 0.5$ weak, $0.5 < |\rho| < 0.8$ moderate, $|\rho| > 0.8$ strong correlation.

In this course all we will do is note trends – a higher level stat course would test to determine whether the correlation is statistically significant. (This will depend in part on both value of r and sample size n).

IMPORTANT: Correlation is not the same thing as causation. No matter how strong the correlation may be, we still cannot say that "X causes Y" or "Y causes X". While it may be so, it might also be true that "both X and Y are the result of some other unknown factor(s)". Consider unemployment vs. inflation, or weight vs. height.

Example F. A manufacturer has collected preliminary data relating number of units produced (X, measured in hundreds) and cost (Y, measured in \$1000's). Data are represented in the following points: (2, 4), (5, 6), (6, 7) and (9, 8). Find the correlation coefficient. *answer*: ≈ 0.9804