Dis 5D: Variance and Covariance

Friday, 17 July 2020 11:17 PM

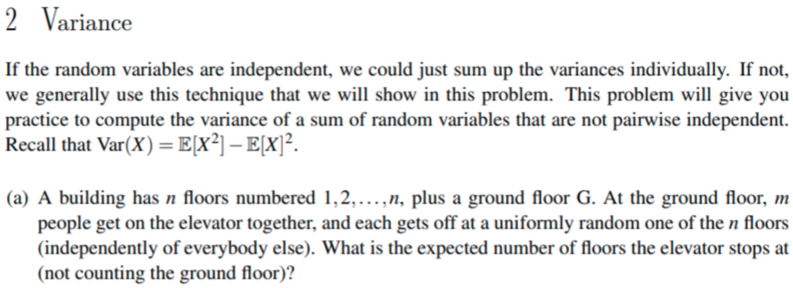
Variance of Random Variables (X)

Intuitively, Var(X) is how much the random variable fluctuates in its values. A random variable that only takes on one value would have no variance.

E.g. Random experiment = flip a coin once; RV = number of coins flipped …

Formally, Var(X) = E[(X – E[X])^2] i.e. the expected value of the square of the difference of a random variable from its mean

Megan writes letters to her n best friends. Because of a hurricane, Megan's parrot mixes up the addresses of Megan's friends. And so, each friend gets a random letter. Let be X the number of friends who get the letter that was meant for them. What's the variance of X?



2 Variance

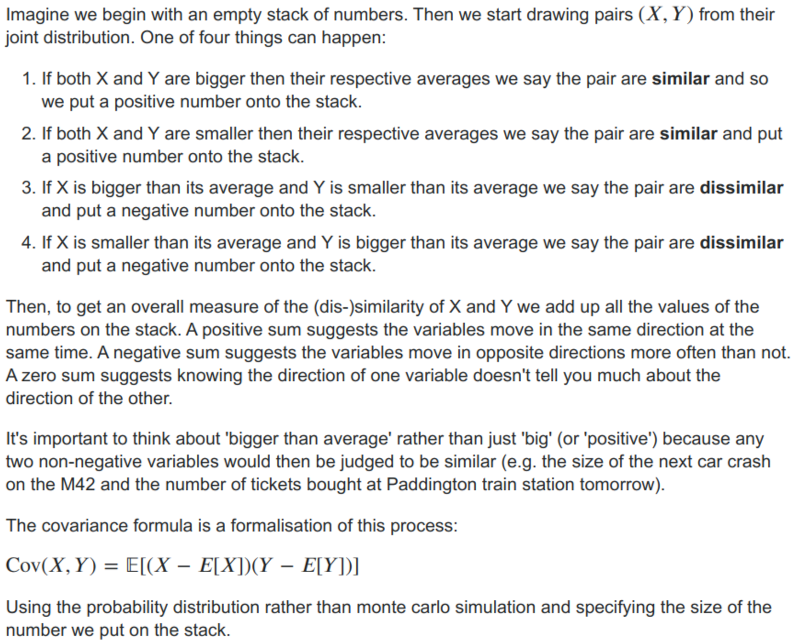
If the random variables are independent, we could just sum up the variances individually. If not, we generally use this technique that we will show in this problem. This problem will give you practice to compute the variance of a sum of random variables that are not pairwise independent. Recall that $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

(a) A building has n floors numbered $1, 2, \ldots, n$, plus a ground floor G. At the ground floor, m people get on the elevator together, and each gets off at a uniformly random one of the n floors (independently of everybody else). What is the expected number of floors the elevator stops at (not counting the ground floor)?

(b) What is the *variance* of the number of floors the elevator *does not* stop at? (In fact, the variance of the number of floors the elevator *does* stop at must be the same (make sure you understand why), but the former is a little easier to compute.)

(c) A group of three friends has n books they would all like to read. Each friend (independently of the other two) picks a random permutation of the books and reads them in that order, one book per week (for n consecutive weeks). Let X be the number of weeks in which all three friends are reading the same book. Compute $\mathrm{Var}(X)$.

Intuition on the definition of the **covariance** (of random variables X, Y)



Imagine we begin with an empty stack of numbers. Then we start drawing pairs (X, Y) from their joint distribution. One of four things can happen:

1. If both X and Y are bigger then their respective averages we say the pair are **similar** and so we put a positive number onto the stack.
2. If both X and Y are smaller then their respective averages we say the pair are **similar** and put a positive number onto the stack.
3. If X is bigger than its average and Y is smaller than its average we say the pair are **dissimilar** and put a negative number onto the stack.
4. If X is smaller than its average and Y is bigger than its average we say the pair are **dissimilar** and put a negative number onto the stack.

Then, to get an overall measure of the (dis-)similarity of X and Y we add up all the values of the numbers on the stack. A positive sum suggests the variables move in the same direction at the same time. A negative sum suggests the variables move in opposite directions more often than not. A zero sum suggests knowing the direction of one variable doesn't tell you much about the direction of the other.

It's important to think about 'bigger than average' rather than just 'big' (or 'positive') because any two non-negative variables would then be judged to be similar (e.g. the size of the next car crash on the M42 and the number of tickets bought at Paddington train station tomorrow).

The covariance formula is a formalisation of this process:

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - E[X])(Y - E[Y])]$$

Using the probability distribution rather than monte carlo simulation and specifying the size of the number we put on the stack.

Source: conjectures (https://stats.stackexchange.com/users/16663/conjectures), Intuition on the definition of the covariance, URL (version: 2014-05-17): https://stats.stackexchange.com/q/99096

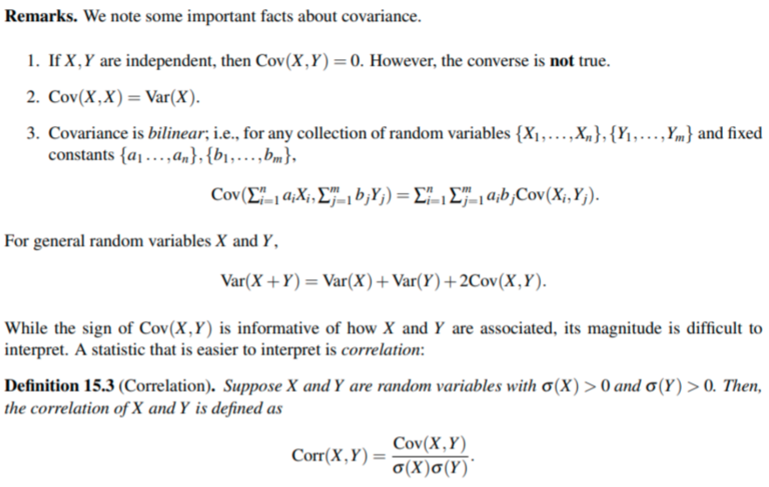
If cov(X, Y) is positive, then X and Y are "similar". Similar means when X increases, Y increases. When X decreases, Y decreases.

E.g.

Random Experiment: take a random student in UC Berkeley

Let X be the RV for the age of the student
Let Y be the RV for the number of exams they've aced since coming to Berkeley

Cov(X, Y) should be ___________________



Remarks. We note some important facts about covariance.

1. If X, Y are independent, then $\mathrm{Cov}(X,Y) = 0$. However, the converse is **not** true.
2. $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.
3. Covariance is *bilinear*; i.e., for any collection of random variables $\{X_1, \ldots, X_n\}, \{Y_1, \ldots, Y_m\}$ and fixed constants $\{a_1 \ldots, a_n\}, \{b_1, \ldots, b_m\}$,

$$\mathrm{Cov}(\textstyle\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \mathrm{Cov}(X_i, Y_j).$$

For general random variables X and Y,

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).$$

While the sign of $\mathrm{Cov}(X,Y)$ is informative of how X and Y are associated, its magnitude is difficult to interpret. A statistic that is easier to interpret is *correlation*:

Definition 15.3 (Correlation)**.** *Suppose X and Y are random variables with $\sigma(X) > 0$ and $\sigma(Y) > 0$. Then, the correlation of X and Y is defined as*

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)}.$$

3 Covariance

We have a bag of 5 red and 5 blue balls. We take two balls uniformly at random from the bag without replacement. Let X_1 and X_2 be indicator random variables for the first and second ball being red. What is $\text{cov}(X_1, X_2)$? Recall that $\text{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$.

1 Ball in Bins

You are throwing k balls into n bins. Let X_i be the number of balls thrown into bin i.

(a) What is $\mathbb{E}[X_i]$?

(b) What is the expected number of empty bins?

(c) Define a collision to occur when two balls land in the same bin (if there are n balls in a bin, count that as $n-1$ collisions). What is the expected number of collisions?