Self-paced Example: Correlated random variables

Module 10

MCIT Online - CIT592 - Professor Val Tannen

This is a segment that contains material meant to be learned *at your own pace*. We are trying to assist you in this endeavor by organizing the material in a manner similar to the way it is outlined in the recorded segments, however with one additional suggestion. When you see the following marker:

we suggest that you stop and make sure you thoroughly understood the material presented so far before you proceed further.

Product of two r.v.'s

Let $X, Y : \Omega \to \mathbb{R}$ be two random variables defined on the same probability space (Ω, Pr). Their **product**, notation XY, is the random variable on the same space defined by:

$$(XY)(w) = X(w) \cdot Y(w) \quad \forall\, w \in \Omega$$

Example. A fair coin is flipped twice. Recall that we denoted by X_H and X_T the random variables that return, respectively, the number of heads and tails observed.

The probability space of two independent flips of a fair coin is uniform and has the outcomes $\{\mathtt{HH}, \mathtt{HT}, \mathtt{TH}, \mathtt{TT}\}$ each with probability $1/4$.

Then $X_H X_T(\mathtt{HH}) = X_H X_T(\mathtt{TT}) = 0$ and $X_H X_T(\mathtt{HT}) = X_H X_T(\mathtt{TH}) = 1$.

So $X_H X_T$ is a Bernoulli r.v. with parameter $1/2$.

Correlated random variables

Proposition $\mathsf{Var}[X + Y] = \mathsf{Var}[X] + \mathsf{Var}[Y]$ iff $\mathsf{E}[XY] = \mathsf{E}[X] \cdot \mathsf{E}[Y]$.

Proof Applying LOE multiple times we have

$$\begin{aligned}
\mathsf{Var}[X+Y] &= \mathsf{E}[(X+Y)^2] - (\mathsf{E}[X+Y])^2 \\
&= \mathsf{E}[X^2 + 2XY + Y^2] - (\mathsf{E}[X] + \mathsf{E}[Y])^2 \\
&= \mathsf{E}[X^2] + 2\mathsf{E}[XY] + \mathsf{E}[Y^2] - (\mathsf{E}[X])^2 - 2(\mathsf{E}[X] \cdot \mathsf{E}[Y]) - (\mathsf{E}[Y])^2 \\
&= \mathsf{Var}[X] + \mathsf{Var}[Y] - 2(\mathsf{E}[XY] - \mathsf{E}[X] \cdot \mathsf{E}[Y])
\end{aligned}$$

This suggests that we *define*: the random variables X and Y are **correlated** when $\mathsf{E}[XY] \neq \mathsf{E}[X] \cdot \mathsf{E}[Y]$. Thus variance distributes over the sum of uncorrelated r.v.'s.

To test yourself, think of a simple example of two specific correlated random variables.

Example. The r.v.'s X_H and Y_H defined earlier are correlated. Indeed, we have calculated in a previous lecture that $\mathsf{E}[X_H] = \mathsf{E}[X_T] = 1$. However, $X_H X_T$ is Bernoulli with parameter $1/2$ so $\mathsf{E}[X_H X_T] = 1/2 \neq 1 = \mathsf{E}[X_H] \cdot \mathsf{E}[X_T]$.

Notice how a higher X_H means a lower X_T. This case is referred to as a **negative correlation**. To give this notion mathematical precision, define **covariance** as:

$$\mathsf{Cov}(X, Y) = \mathsf{E}[XY] - \mathsf{E}[X]\mathsf{E}[Y]$$

Therefore, two random variables are uncorrelated if and only if their covariance is 0. By contrast, in our example, $\mathsf{Cov}(X_H, X_T) = \frac{1}{2} - 1 = -\frac{1}{2}$, a negative number. If the covariance had been positive, then we would have had a **positive correlation** between the variables.

Correlated versus independent random variables

Recall that two random variables are independent when for any x in $\mathrm{Val}(X)$ and any y in $\mathrm{Val}(Y)$,

$$\Pr[(X = x) \cap (Y = y)] = \Pr[X = x] \cdot \Pr[Y = y]$$

Proposition If two random variables are independent then they are uncorrelated.

Proof Suppose the random variables X and Y are independent. By the definition of product of random variables,

$$\mathsf{E}[XY] = \sum_{x \in \mathrm{Val}(X)} \sum_{y \in \mathrm{Val}(Y)} \Pr[(X = x) \cap (Y = y)] \cdot x \cdot y$$

By the definition of independent random variables then,

$$\mathsf{E}[XY] = \sum_{x \in \mathrm{Val}(X)} \sum_{y \in \mathrm{Val}(Y)} \Pr[X = x] \cdot \Pr[Y = y] \cdot x \cdot y$$

By rearranging the terms that depend on x and on y:

$$\mathsf{E}[XY] = \sum_{x \in \mathrm{Val}(X)} \Pr[X = x] \cdot x \cdot \sum_{y \in \mathrm{Val}(Y)} \Pr[Y = y] \cdot y$$

Therefore, by definition of expectation, $\mathsf{E}[XY] = \mathsf{E}[X] \cdot \mathsf{E}[Y]$.

Now, can you think of two random variables that are uncorrelated but not independent? This is possible!

Example Let X be a random variable that takes the values -1, 0, and 1 each with probability $\frac{1}{3}$. Hence $\mathsf{E}[X] = 0$. Let $Y = X^2$ be our second random variable. Note that X^n whenever n is odd has the exact same distribution as X. Computing the covariance:

$$\mathsf{Cov}(X, X^2) = \mathsf{E}[X^3] - \mathsf{E}[X] \cdot \mathsf{E}[X^2] = 0 - 0 \cdot \mathsf{E}[X^2] = 0$$

We have shown X and Y are uncorrelated, but we have to show they are not independent. Consider two events: $(X = 0)$ and $(Y = 1)$. Note that $\Pr[X = 0 \text{ and } Y = 1] = 0$. However,

$$\Pr[X = 0] \cdot \Pr[Y = 1] = \frac{1}{3} \cdot \frac{2}{3} \neq 0$$

Hence the random variables are dependent. This result might seem counterintuitive, but it is explained by the fact that correlation is a measure of *linear* dependence. In our example, the overall covariance between X and Y may be 0, but this is only because the negative contribution from $X = -1$ exactly cancels the positive contribution from $X = 1$.

More than two random variables

We have shown above that if two random variables are uncorrelated then variance distributed over their sum. That is, $\mathsf{Var}[X + Y] = \mathsf{Var}[X] + \mathsf{Var}[Y]$ when (in fact, iff) $\mathsf{E}[XY] = \mathsf{E}[X] \cdot \mathsf{E}[Y]$. Now we will extend this to more than two random variables.

Proposition If $X_1, \ldots, X_n$ are *pairwise uncorrelated*, then:

$$\mathsf{Var}[X_1 + \cdots + X_n] \;=\; \mathsf{Var}[X_1] + \cdots + \mathsf{Var}[X_n]$$

Proof To keep the notation lighter we just sketch this for three random variables. Using linearity of expectation, you can see that (verify!):

$$\begin{aligned} \mathsf{Var}[\,X + Y + Z\,] &= \mathsf{Var}[X] + \mathsf{Var}[Y] + \mathsf{Var}[Z] + \\ &= 2(\mathsf{E}[XY] - \mathsf{E}[X]\mathsf{E}[Y]) + 2(\mathsf{E}[YZ] - \mathsf{E}[Y]\mathsf{E}[Z]) + 2(\mathsf{E}[ZX] - \mathsf{E}[Z]\mathsf{E}[X]) \end{aligned}$$

Since X, Y, Z are pairwise uncorrelated the last three terms in the sum are 0 and the identity follows.