Dr. Z.'s Probability Lecture 19 Handout: Covariance, Variance of Sums, and Correlations

By Doron Zeilberger

Version of Nov. 21, 2017 (Thanks to Zuzanna Kasper)

Obvious but Important Fact: If X and Y are **independent** random variables, then

$$E[XY] = E[X]\,E[Y] \quad ,$$

in other words the expectation of the product equals the product of the expectations.

More generally, for any functions $g(X)$ and $h(Y)$ of X and Y, if If X and Y are **independent** random variables, then

$$E[g(X)h(Y)] = E[g(X)]\,E[h(Y)] \quad ,$$

Note: The general version follows immediately from the special version, since the random variables $g(X)$ and $h(Y)$ are independent.

Another Note: Sometimes it happens that $E[XY] = E[X]E[Y]$ even when X and Y are not independent. Then X and Y are called *uncorrelated*. So independence implies no correlation (but not vice versa).

Important Definition (Covariance) If X and Y are random variables on the same probability space, the **covariance** of X and Y, denoted by $Cov(X,Y)$, is defined by

$$Cov(X,Y) \;=\; E[\,(X - E[X])\,(Y - E[Y])] \quad .$$

Note: This formula is pretty intimidating. It is more useful to say: If X and Y are random variables with means μ_1 and μ_2 respectively, then

$$Cov(X,Y) \;=\; E[\,(X - \mu_1)\,(Y - \mu_2)] \quad .$$

Useful formula

$$Cov(X,Y) = E[XY] \;-\; E[X]\,E[Y] \quad .$$

Problem 19.1: Suppose that

- Alex is 200cm -tall, and weighs 70kgs ;
- Bob is 180cm -tall, and weighs 80kgs ;
- Charlie is 190cm -tall, and weighs 100kgs ;

• Dave is 205cm -tall, and weighs 95kgs ;

At each game, the coach picks one of them to be the captain of the team. But he has favorites. The probability that he will pick Alex, Bob, Charlie, or Dave are $0.1, 0.2, 0.3, 0.4$ respectively.

Let X and Y be the random variables "height of captain" and "weight of captain". Find the covariance between X and Y.

Sol. to 19.1: First we need $E[X]$ and $E[Y]$.

$$E[X] \,=\, 0.1 \cdot 200 \,+\, 0.2 \cdot 180 + \, 0.3 \cdot 190 + \, 0.4 \cdot 205 \,=\, 195 \quad .$$

$$E[Y] \,=\, 0.1 \cdot 70 \,+\, 0.2 \cdot 80 + \, 0.3 \cdot 100 + \, 0.4 \cdot 95 \,=\, 91 \quad .$$

To compute $Cov(X, Y)$ let's do it in two ways.

First Way (Using the formula)

$$E[XY] \,=\, 0.1 \cdot 200 \cdot 70 \,+\, 0.2 \cdot 180 \cdot 80 + \, 0.3 \cdot 190 \cdot 100 + \, 0.4 \cdot 205 \cdot 95 \,=\, 17770.0 \quad .$$

Hence

$$Cov(X, Y) = E[XY] - E[X]\, E[Y] \,=\, 17770 - 195 \cdot 91 = 25 \quad .$$

Second Way (Using the definition)

$$E[(X - E[X])(Y - E[Y])] \,=$$

$$0.1{\cdot}(200{-}195){\cdot}(70{-}91) + 0.2{\cdot}(180{-}195){\cdot}(80{-}91) + 0.3{\cdot}(190{-}195){\cdot}(100{-}91) + 0.4{\cdot}(205{-}195){\cdot}(95{-}91) \,= 25 \quad .$$

Ans. to 19.1: The covariance of the height and weight of the chosen captain, $Cov(X, Y)$, equals 25.

Obvious but Important Formulas

$$Cov(X, Y) = Cov(Y, X) \quad ,$$

$$Cov(X, X) = Var(X) \quad ,$$

$$Cov(aX, Y) = aCov(X, Y) \quad ,$$

$$Cov\left(\sum_{i=1}^{n} X_i \,,\, \sum_{j=1}^{n} Y_j\right) \,=\, \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(X_i, Y_j) \quad .$$

Important Special Case of the Last Formula

$$Var\left(\sum_{i=1}^{n} X_i\right) \,=\, \sum_{i=1}^{n} Var(X_i) \,+\, 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} Cov(X_i, X_j)$$

Note: In particular, $Var(X+Y) = Var(X) + Var(Y) + 2\,Cov(X,Y)$.

Very important Consequence: If $X_1, \ldots, X_n$ are **pairwise independent** then

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) \quad .$$

Important Definition: The **Correlation** between two random variables X and Y, denoted by $\rho(X,Y)$, is the covariance divided by the product of the standard deviations of X and Y. In symbols:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \quad .$$

Note: The correlation is always between -1 and 1. It is a measure of how "related" X and Y are.

Problem 19.2: In a certain family of four children, the scores (out of 100) for English and Math are as follows

Abe: English: 90 ; Math: 80

Bob: English: 95 ; Math: 40

Charlie: English: 50 ; Math: 70

Dave: English: 85 ; Math: 98

What is The correlation between the English scores and the Math scores?

Sol. to 19.2: In such a problem the probability of each person is the same, in this case $\frac{1}{4}$. let X be the English score and Y be the Math score.

$$E[X] = \frac{1}{4}(90+95+50+85) = 80 \quad ,$$

$$E[Y] = \frac{1}{4}(80+40+70+98) = 72 \quad .$$

Next

$$Var(X) = \frac{1}{4}\left((90-80)^2 + (95-80)^2 + (50-80)^2 + (85-80)^2\right) = \frac{625}{2} = 312.5 \quad ,$$

$$Var(Y) = \frac{1}{4}\left((80-72)^2 + (40-72)^2 + (70-72)^2 + (98-72)^2\right) = 442 \quad ,$$

$$Cov(X,Y)$$

$$= \frac{1}{4}\Big((90-80)\cdot(80-72) + (95-80)\cdot(70-72) + (50-80)\dot{(}70-72) + (85-80)\cdot(98-72)\Big)$$

$$= -52.5 \quad .$$

Finally

$$\rho(X,Y) \;=\; \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

$$= \frac{-52.5}{\sqrt{312.5\cdot 442}} = -0.1412612867\ldots \quad .$$

Ans. to 19.2: The correlation between the English and Math scores in that family is $-0.1412612867\ldots$. So it is a little negative.

Problem 19.3: If $Var(X) = 5$, $Var(Y) = 7$, and $Var(X+Y) = 14$, find

(i) $Var(2X+3Y)$;

(ii) $Var(X-Y)$.

Sol. to 19.3: **First** we need to find $Cov(X,Y)$, using

$$Var(X+Y) \;=\; Var(X) \;+\; Var(Y) \;+\; 2\,Cov(X,Y) \quad .$$

So

$$14 \;=\; 5 \;+\; 7 \;+\; 2\,Cov(X,Y) \quad ,$$

hence $Cov(X,Y) = 1$.

Now we can answer the questions. Recall that

$$Var(aX) \;=\; a^2\,Var(X) \quad ; \quad Cov(aX,bY) \;=\; ab\,Cov(X,Y) \quad .$$

Sol. of (i):

$$Var(2X+3Y) \;=\; Var(2X)+Var(3Y)+2\,Cov(2X,3Y) \;=\; 2^2\cdot Var(X)+3^2\cdot Var(Y)+2\,Cov(2X,3Y)$$

$$= \;4\,Var(X) \;+\; 9\,Var(Y) \;+\; 2\cdot 2\cdot 3\,Cov(X,Y)$$

$$= \;4\cdot 5 + 9\cdot 7 + 12\cdot 1 = 20 + 63 + 12 = 95 \quad .$$

Ans. to 19.3(i): 95.

Sol. of (ii):

$$Var(X-Y) \;=\; Var(X)+Var(-Y)+2\,Cov(X,-Y) \;=\; Var(X)+(-1)^2 Var(Y)+2\cdot(-1)\,Cov(X,Y)$$

$$= 5 + 7 - 2\cdot 1 = 10 \quad .$$

Ans. to 19.3(ii): 10.

Problem 19.4 : An insurance policy pays a total medical benefit consisting of two parts for each claim. Let X represent the part of the benefit that is paid to the surgeon, and let Y represent the

part that is paid to the hospital. The variances of X is 1000, the variance of Y is 4000, and the variance of the total, $X+Y$, is 6000.

Due to increasing medical costs the company that issues the policy decides to increase X by a finite amount of 300 per claim, and to increase Y by 20% per claim.

Calculate the variance of the total benefit after these revisions are made.

Sol. to 19.4: First we need to find $Cov(X,Y)$. Since

$$Var(X+Y) = Var(X) + Var(Y) + 2\,Cov(X,Y) \quad ,$$

we have

$$6000 = 1000 + 4000 + 2\,Cov(X,Y) \quad ,$$

Hence $Cov(X,Y) = 500$.

The **new** total benefit is $X + 300 + 1.2Y$, so we have (recall that adding a constant to X does not change its variance)

$$Var(X+300+1.2Y) = Var(X+1.2Y) = Var(X) + Var(1.2Y) + 2\,Cov(X,1.2Y) =$$

$$= Var(X) + 1.2^2 \cdot Var(Y) + 2 \cdot 1.2\,Cov(X,Y) = 1000 + 1.2^2 \cdot 4000 + 2 \cdot 1.2 \cdot 500 = 7960 \quad .$$

Ans. to 19.4: The variance of the total benefit after the revisions are made is 7960.

Problem 19.5: Let X and Y be the number of hours that a randomly selected person listens to classical music and Jazz, respectively, during a one-month period. The following information is known about X and Y.

$$E[X] = 10 \quad , \quad E[Y] = 20 \quad , \quad Var(X) = 10 \quad , \quad Var(Y) = 30 \quad , \quad Cov(X,Y) = 12 \quad .$$

Four hundred people are randomly selected and observed for this month. Let T be the total number of hours that these four hundred people listen to Classical and Jazz this month.

Approximate the value of $P(T < 12080)$.

Note: By the Central Limit Theorem (coming up in Lecture 23) if $T_1, \ldots, T_n$ are independent and each with the same distribution with mean μ and variance σ^2 then the distribution

$$\frac{T_1 + \ldots + T_n - n\mu}{\sigma\sqrt{n}}$$

is close to the Standard Normal Distribution.

Sol. Let T_i be the total listening time for *one* person. $T_i = X + Y$. Then

$$Var(T_i) \;=\; Var(X) \;+\; Var(Y) \;+\; 2\,Cov(X,Y) \;=\; 10 + 30 + 2 \cdot 12 \;=\; 64 \quad .$$

Also $E[T_i] = E[X] + E[Y] = 10 + 20 = 30$. So $\mu = 30$ and $\sigma = 8$, for each i. Let

$$Z = \frac{T_1 + \ldots + T_{400} - 400\mu}{\sigma\sqrt{400}} = \frac{T - 400 \cdot 30}{8 \cdot 20} = \frac{T - 12000}{160}$$

So

$$P(T < 12080) \; = \; P(\frac{T - 12000}{160} < \frac{12080 - 12000}{160}) = P(\frac{T - 12000}{160} < 0.5) \quad .$$

Since $\frac{T-12000}{160}$ is approximately the standard normal distribution this is approximately $\Phi(0.5) = 0.6915$.

Ans. to 19.5: The probability that $P(T < 12080)$ is approximately %69.15.

Note: This problem really belongs to Lecture 23, but a crucial part is to find the variance of T_i, so I put it in this lecture. Also the Central Limit Theorem is an important theorem so it is good to be already exposed to it earlier.