Lecture Notes EMSE 4765: DATA ANALYSIS - Probability Review

Chapter 10: Covariance and Correlation

Version: 1/19/2021

Text Book: A Modern Introduction to Probability and Statistics, Understanding Why and How

By: F.M. Dekking. C. Kraaikamp, H.P.Lopuhaä and L.E. Meester

10 Covariance and Correlation

10.1 Expectation and joint distributions…

- In this chapter we see how **the joint distribution of two (or more) random variables, say X and Y,** is used to compute the expectation of a combination of these random variables, **e.g. $g(X,Y)$. That, is how to compute $E[\,g(X,Y)]$.**

- More specifically, we discuss **the expectation $E[X+Y]$ and $E[X \cdot Y]$ and the variance of a sum of random variables $V[X+Y]$** as important special cases and introduce in the process the notions of **covariance $Cov(X,Y)$ and correlation $\rho(X,Y)$**, which express the way two random variables influence each other.

- Recall that for the situation where **random variables <u>do not provide information to one another</u> the notion of statistical independence** for random variables applies. **In that case, both the covariance $Cov(X,Y)$ and correlation $\rho(X,Y)$ turn outs to be equal to zero.** The converse, however, is not true!

10 Covariance and Correlation

10.1 Expectation and joint distributions...

Two-dimensional change-of-variable formula. Let X and Y be random variables, and let $\boldsymbol{g : \mathbb{R}^2 \longrightarrow \mathbb{R}}$ **be a function**. If X and Y are discrete rv's with outcomes $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$, respectively, with **joint pmf** $\boldsymbol{Pr(X = a_i, Y = b_j)}$ and $\boldsymbol{Z = g(X, Y)}$ then :

$$E[Z] = E[g(X,Y)] = \sum_i \sum_j g(a_i, b_j) Pr(X = a_i, Y = b_j)$$

Two-dimensional change-of-variable formula. Let X and Y be rv's and let $g : \mathbb{R}^2 \longrightarrow \mathbb{R}$ be a function. If X and Y are continuous rv's with **joint probability density function** $\boldsymbol{f_{X,Y}(\cdot\,,\,\cdot)}$ and $\boldsymbol{Z = g(X, Y)}$ then :

$$E[Z] = E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) dx dy$$

Example : (Table 10.1) Given Joint Probabilities $P(X = a, Y = b)$:

	a		
b	0	1	2
0	0	1/4	0
1	1/4	0	1/4
2	0	1/4	0

Exercise: Compute $E[X+Y]$ for the random variables with the joint distribution given in Table 10.1.

Answer: Defining $Z = g(X,Y) = X + Y$ we have

$$
\begin{aligned}
E[Z] = E[X+Y] &= (0+0)\cdot 0 + (0+1)\cdot \frac{1}{4} + (0+2)\cdot 0 \; + (1+0)\cdot \frac{1}{4} \\
&+ (1+1)\cdot 0 + (1+2)\cdot \frac{1}{4} + (2+0)\cdot 0 + (2+1)\cdot \frac{1}{4} \\
&+ (2+2)\cdot 0 = 0 + \frac{1}{4} + 0 + \frac{1}{4} + 0 + \frac{3}{4} + 0 + \frac{3}{4} + 0 = 2.
\end{aligned}
$$

10 Covariance and Correlation

10.1 Expectation and joint distributions...

Exercise: Determine **the marginal distributions** of X and Y in the example above and evaluate $E[X]$, $E[Y]$. **Answer:** Summing columns we have:

$$P(X=0)=P(X=2)=\frac{1}{4},\ P(X=1)=\frac{1}{2}\Rightarrow E[X]=0\cdot\frac{1}{4}+1\cdot\frac{1}{2}+2\cdot\frac{1}{4}=1.$$

Summing rows we have:

$$P(Y=0)=P(Y=2)=\frac{1}{4},\ P(Y=1)=\frac{1}{2}\Rightarrow E[Y]=0\cdot\frac{1}{4}+1\cdot\frac{1}{2}+2\cdot\frac{1}{4}=1.$$

Hence:

$$E[X+Y]=E[X]+E[Y]$$

Is this true in General? Yes, in fact:

Linearity of Expectations. For all numbers r, s, t and rv's X and Y we have

$$E[rX+sY+t]=rE[X]+sE[Y]+t$$

10 Covariance and Correlation

10.1 Expectation and joint distributions...

More generally, for random variables $X_1, \ldots, X_n$ and numbers $s_1, \ldots, s_n$ and t,

$$E\Big[\sum_{i=1}^{n} s_i X_i + t\Big] = E\Big[s_1 X_1 + \ldots + s_n X_n + t\Big]$$

$$= s_1 E[X_1] + \ldots + s_n E[X_n] + t = \sum_{i=1}^{n} s_i E[X_i] + t$$

Exercise: **You are tossing an unfair coin n times** with $P(H) = p$. How many times do you expect to see "heads"? Defining :

$$Z \equiv \text{Number of times you see } H \text{ in a series of } n \text{ coin-tosses, } Z \sim Bin(n, p)$$

We need to calculate:

$$E[Z] = \sum_{k=1}^{n} k \binom{n}{k} p^k (1-p)^{n-k}, \textbf{ Somewhat challenging, no?}$$

Answer: Defining $R_i = 1$ when "the i-th coin toss is H" and 0 otherwise, then:

$$P(R_i = 1) = p,\ P(R_i = 0) = 1 - p,\ E[R_i] = p \cdot 1 + (1 - p) \cdot 0 = p.$$

Thus, $R_i \sim Ber(p)$ and we have:

$$Z = \sum_{i=1}^{n} R_i \Rightarrow E\Big[\sum_{i=1}^{n} R_i\Big] = \sum_{i=1}^{n} E[R_i] = \sum_{i=1}^{n} p = n \cdot p.$$

More than two random variables: This a generalization of the two variable case:

Change-of-variable formula. For example $g : \mathbb{R}^n \to \mathbb{R}$ and $X_1, \ldots X_n$ are continuous rv's with joint density function $f(\cdot)$ and $Z = g(X_1, \ldots, X_n)$ then we have for $E[Z]$:

$$E[g(X_1, \ldots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

10 Covariance and Correlation

10.1 Variance and joint distributions...

We have seen $\boldsymbol{E[X+Y]=E[X]+E[Y]}$, What about $\boldsymbol{Var(X+Y)}$?

$$Var(X+Y)=E\Big[(X+Y-\boldsymbol{E[X+Y]})^2\Big]=E\Big[(X+Y-(\boldsymbol{E[X]+E[Y]}))^2\Big]$$

$$=E\Big[(X-\boldsymbol{E[X]}+Y-\boldsymbol{E[Y]})^2\Big]=E\Big[(\boldsymbol{\{}X-\boldsymbol{E[X]}\boldsymbol{\}}+\boldsymbol{\{}Y-\boldsymbol{E[Y]}\boldsymbol{\}})^2\Big]$$

$$=E\Big[(X-E[X])^2+(Y-E[Y])^2+2(X-E[X])(Y-E[Y])\Big]$$

$$=E\Big[(X-E[X])^2\Big]+E\Big[(Y-E[Y])^2\Big]+$$

$$2E\Big[(X-E[X])(Y-E[Y])\Big]$$

$$=Var(X)+Var(Y)+2E\Big[(X-E[X])(Y-E[Y])\Big]$$

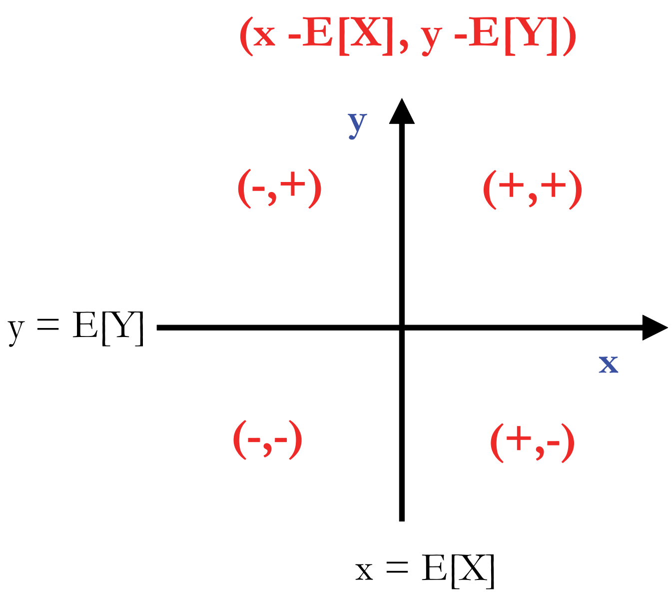
$$=Var(X)+Var(Y)+2Cov(X,Y)$$

So how does $\boldsymbol{Var(X+Y)}$ relate to $\boldsymbol{Var(X)}$ and $\boldsymbol{Var(Y)}$?

Definition: Given two rv's X and Y the covariance of X and Y is defined as:

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

when $\boldsymbol{Cov(X,Y) = 0}$ the rv's X and Y are called **uncorrelated.**



Positive Dependence: **High** values of X tend to go with **high** values of Y and **small** values of X go with **small** value of Y $\Rightarrow$ $\boldsymbol{Cov(X,Y) > 0}$.

Negative Dependence: **High** values of X tend to go with **small** values of Y and **small** values of X go with **high** value of Y $\Rightarrow$ $\boldsymbol{Cov(X,Y) < 0}$.

10 Covariance and Correlation

10.2 Covariance and Independence...

An alternative expression for the covariance: Given two rv's X and Y we have:

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

- **X, Y independent rv's:** Without loss of generality assume that X and Y are discrete r.v's :

$$E[X \cdot Y] = \sum_i \sum_j a_i b_j P(X = a_i, Y = b_j) = \sum_i \sum_j a_i b_j P(X = a_i) P(Y = b_j)$$

$$= \sum_i a_i P(X = a_i) \left[\sum_j b_j P(Y = b_j) \right] = \sum_i a_i P(X = a_i) \cdot E[Y]$$

$$= E[Y] \cdot \sum_i a_i P(X = a_i) = E[X] \cdot E[Y].$$

Independent versus uncorrelated: If two rv's X and Y are independent, then X and Y are uncorrelated, i.e. $Cov(X,Y) = 0$.

When $Cov(X,Y) = 0$ does that imply that X, Y independent rv's?

Example : Given Joint Probabilities

$P(X = a, Y = b)$:

b \ a	0	1	2
0	0	1/4	0
1	1/4	0	1/4
2	0	1/4	0

Recall $E[X] = E[Y] = 1$ and

$$E[X \cdot Y] =$$
$$(0 \cdot 0) \cdot 0 + (0 \cdot 1) \cdot \tfrac{1}{4} + (0 \cdot 2) \cdot 0 +$$
$$(1 \cdot 0) \cdot \tfrac{1}{4} + (1 \cdot 1) \cdot 0 + (1 \cdot 2) \cdot \tfrac{1}{4} +$$
$$(2 \cdot 0) \cdot 0 + (2 \cdot 1) \cdot \tfrac{1}{4} + (2 \cdot 2) \cdot 0 +$$
$$= 1$$

10 Covariance and Correlation

10.1 Covariance and independence...

Hence, in this particular example:

$$E[X] \cdot E[Y] = 1 \cdot 1 = E[X \cdot Y] = 1 \Leftrightarrow Cov(X, Y) = 0$$

Are the random variables X and Y independent?

- **Answer:** When you know the value of the random variable X equals 0 or 2, you know the value of the random variable Y. Hence, the answer is: **No!**

- **In general:** When information of one random variable tells you something about the other random variables, those random variable are **Statistically Dependent.**

- X, Y independent random variable $\Rightarrow E[X \cdot Y] = E[X] \cdot E[Y]$.

- The converse is not necessarily true, i.e. **if X and Y are uncorrelated, they need not be statistically independent**

Variance of the sum: Given two rv's X and Y then always

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

If X and Y are **independent or uncorrelated** then:

$$Var(X+Y) = Var(X) + Var(Y).$$

- The variance of a sum of uncorrelated rv's is equal to their sum of the variances. That is: if $X_1, \ldots, X_n$ are uncorrelated then

$$Var(X_1 + \ldots + X_n) = Var(X_1) + \ldots + Var(X_n).$$

Example: Recall tossing a unfair coin n times. Then $Z \equiv$ "# times you see "heads", $Z \sim Bin(n,p)$ and $Z \equiv \sum_{i=1}^{n} R_i,\ R_i \sim Ber(p)$, R_i independent rv's.

$$\begin{aligned} Var(R_i) &= \boldsymbol{E[(R_i)^2]} - \boldsymbol{(E[R_i])^2} = \boldsymbol{0^2 \cdot (1-p) + 1^2 \cdot p} - \boldsymbol{p^2} \\ &= \boldsymbol{p} - \boldsymbol{p^2} = p(1-p). \end{aligned}$$

Now using $Z = \sum_{i=1}^{n} R_i$ and the fact that the rv's R_i are independent rv's we have:

$$Var(Z) = Var\left(\sum_{i=1}^{n} R_i\right) = \sum_{i=1}^{n} Var(R_i) = \sum_{i=1}^{n} p(1-p) = np(1-p).$$

- **Summarizing:**

$$\boldsymbol{R_i \sim Ber(p) \Rightarrow E[R] = p,\ Var(R) = p(1-p),\ R_i \text{ ind. rv's}}$$

$$\boldsymbol{Z = \sum_{i=1}^{n} R_i \Rightarrow Z \sim Bin(n,p), E[Z] = np, Var(Z) = np(1-p)}$$

10 Covariance and Correlation

10.3 Covariance and joint distributions…

- Recall covariance is a measure of positive of negative dependence, i.e. it measures how two rv's X and Y influence one another.

Example: Suppose L and W are rv's modeling **the height and weight** of Dutch citizens **in inches (in) and kilograms (kg)**, respectively. Suppose, **one prefers length in centimeters (cm)**. Defining $L^* = \text{Length in cm} \Rightarrow L^* = \mathbf{2.53}L$.

$$\begin{aligned}
Cov(L^*, W) &= E[L^* \cdot W] - E[L^*] \cdot E[W] \\
&= E[(\mathbf{2.53}L) \cdot W] - E[\mathbf{2.53}L] \cdot E[W] \\
&= \mathbf{2.53}E[L \cdot W] - \mathbf{2.53}E[L] \cdot E[W] \\
&= \mathbf{2.53}(E[L \cdot W] - E[L] \cdot E[W]) \\
&= \mathbf{2.53}Cov(L, W)
\end{aligned}$$

Conclusion: The value of Covariance is affected by a change of scale!

Is the dependence or influence between height and weight any different?

- In general we have for two rv's X and Y and numbers a_x, b_x, a_y and b_y and defining $V \equiv a_x X + b_x,\ W \equiv a_y Y + b_y$:

$$\begin{aligned}
\boldsymbol{V - E[V]} &= a_x X + b_x - E[a_x X + b_x] = a_x X + b_x - a_x E[X] - b_x \\
&= a_x X - a_x E[X] = \boldsymbol{a_x (X - E[X])} \\
\boldsymbol{W - E[W]} &= a_y Y + b_y - E[a_y Y + b_y] = a_y Y + b_y - a_y E[Y] - b_y \\
&= \ = a_y Y - a_y E[Y] = \boldsymbol{a_y (Y - E[Y])} \\
Cov(V, W) &= Cov(a_x X + b_x, a_y Y + b_y) = E[(\boldsymbol{V - E[V]})(\boldsymbol{W - E[W]})] \\
&= E[\boldsymbol{a_x (X - E[X]) a_y (Y - E[Y])}] \\
&= \boldsymbol{a_x a_y} E[\boldsymbol{(X - E[X])(Y - E[Y])}] = \boldsymbol{a_x a_y} Cov(X, Y)
\end{aligned}$$

Covariance under a change-of-units: Given two rv's X and Y then for all numbers a_x, b_x, a_y and b_y :

$$Cov(a_x X + b_x, a_y Y + b_y) = a_x a_y Cov(X, Y).$$

Correlation: Given two rv's X and Y. **The correlation coefficient** $\boldsymbol{\rho(X,Y)} \equiv 0$ when $Var(X) = 0$ or $Var(Y) = 0$ and otherwise:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}} \Rightarrow \rho(X,X) = \frac{Var(X)}{\sqrt{Var(X) \cdot Var(X)}} = 1$$

Correlation under a change-of-units: Recall that $\boldsymbol{\sqrt{x^2} \equiv |x|}$. Given two rv's X and Y, then for all numbers a_x, b_x, a_y and b_y where $a_x, a_y \neq 0$:

$$\rho(a_x X + b_x, a_y Y + b_y) = \frac{a_x \cdot a_y}{\sqrt{a_x^2 \cdot a_y^2}} \rho(X,Y) = \begin{cases} -\rho(X,Y), & a_x \cdot a_y < 0, \\ \rho(X,Y), & a_x \cdot a_y > 0. \end{cases}$$

- When rv's are identical, i.e. $X = Y$, then $\rho(X, Y) = 1$. When rv's only differ in their sign, i.e. $X = -Y$, then $\rho(X, Y) = -1$. **In general, we have**

$$-1 \leq \rho(X, Y) \leq 1, \text{ provided } X \text{ or } Y \text{ are not constant.}$$

Proof: Since we know that the variance of a random variable is non-negative:

$$Var\left(\frac{X}{\sqrt{Var(X)}} + \frac{Y}{\sqrt{Var(Y)}}\right) = Var(a_x X + a_y Y) \geq 0, \text{ where}$$

$$a_x = \frac{1}{\sqrt{Var(X)}}, a_y = \frac{1}{\sqrt{Var(Y)}} \qquad (1)$$

We also know that:

$$Var(a_x X + a_y Y) = Var\Big(a_x X\Big) + Var\Big(a_y Y\Big) + 2Cov\Big(a_x X, a_y Y\Big) \qquad (2)$$

From (1) and (2) it now follows that:

$$
\begin{aligned}
Var(a_x X + a_y Y) &= (\boldsymbol{a_x})^2 Var\Big(X\Big) + (\boldsymbol{a_y})^2 Var\Big(X\Big) + 2\boldsymbol{a_x}\boldsymbol{a_y} Cov\Big(X, Y\Big) \\
&= \frac{Var(X)}{\boldsymbol{Var(X)}} + \frac{Var(Y)}{\boldsymbol{Var(Y)}} + 2\frac{Cov(X,Y)}{\sqrt{\boldsymbol{Var(X)}\boldsymbol{Var(Y)}}} \\
&= 2 + 2\rho(X,Y) \geq 0 \Leftrightarrow \rho(X,Y) \geq -1
\end{aligned}
$$

Using same argument but replacing X by $-X$ one obtains $\rho(X,Y) \leq 1$. $\square$

Suppose $Y = X$, we then have:

$$
\rho(X,X) = \frac{Cov(X,X)}{\sqrt{Var(X)\cdot Var(X)}} = \frac{E\Big[(X - E[X])(X - E[X])\Big]}{\sqrt{Var(X)\cdot Var(X)}}
$$

$$
= \frac{E[(X - E[X])^2]}{\sqrt{[Var(X)]^2}} = \frac{Var(X)}{Var(X)} = 1
$$

$$\rho(X,-X)=\frac{Cov(X,-X)}{\sqrt{Var(X)\cdot Var(-X)}}=\frac{-Cov(X,X)}{\sqrt{Var(X)\cdot Var(X)}}=-\frac{Var(X)}{Var(X)}=-1$$

Correlation under linear realionship: Recalling $\sqrt{x^2}=|x|$. Given two rv's X and $Y=aX+b$ such that $a\neq 0$:

$$\rho(X,aX+b)=\frac{a}{\sqrt{a^2}}\rho(X,X)=\begin{cases}-1, & a<0,\\ 1, & a>0.\end{cases}$$

Conclusion: $\rho(X,Y)$ is called **a measure of linear dependence** since:

$$Y=aX+b,\ a>0\Leftrightarrow\rho(X,X)=1$$

$$Y=aX+b,\ a<0\Leftrightarrow\rho(X,X)=-1$$