P5 - Multiple random variables

STAT 401 (Engineering) - Iowa State University

January 22, 2018

Multiple discrete random variables

Real problems very seldom concern a single random variable. As soon as more than 1 variable is involved it is not sufficient to think of modeling them only individually - their joint behavior is important.

Definition

If X and Y are two discrete variables. Their joint probability mass function is defined as

$$p_{X,Y}(x,y) = P(X = x \cap Y = y) = P(X = x, Y = y).$$

Example

A box contains 5 unmarked PowerPC G4 processors of different speeds:

#	speed
2	400 mHz
1	450 mHz
2	500 mHz

Select two processors out of the box (without replacement) and let

- X be speed of the first selected processor
- Y be speed of the second selected processor

Example (cont.)

Enumerate all the equally probable events:

		1st processor (X)				
	Ω	400_1	400_2	450	500_1	500_2
2nd processor (Y)	400_1	-	x	x	x	x
	400_2	x	-	x	x	x
	450	x	x	-	x	x
	500_1	x	x	x	-	x
	500_2	x	x	x	x	-

Probability mass function:

		1st processor (X)		
	mHz	400	450	500
2nd processor (Y)	400	2/20	2/20	4/20
	450	2/20	0/20	2/20
	500	4/20	2/20	2/20

Example (cont.)

What is the probability that $X = Y$?

$$\begin{aligned} P(X = Y) &= p_{X,Y}(400, 400) + p_{X,Y}(450, 450) + p_{X,Y}(500, 500) \\ &= 2/20 + 0/20 + 2/20 = 4/20 = 0.2 \end{aligned}$$

What is the probability that $X > Y$?

$$\begin{aligned} P(X > Y) &= p_{X,Y}(450, 400) + p_{X,Y}(500, 400) + p_{X,Y}(500, 450) \\ &= 2/20 + 4/20 + 2/20 = 8/20 = 0.4 \end{aligned}$$

Marginal distribution

Definition

For discrete random variables X and Y, the marginal probability mass functions are

$$\begin{aligned} p_X(x) &= \textstyle\sum_y p_{X,Y}(x,y) \\ p_Y(y) &= \textstyle\sum_x p_{X,Y}(x,y) \end{aligned}$$

In the CPU example, we have

x	400	450	500 (mHz)
$p_X(x)$	0.4	0.2	0.4

y	400	450	500 (mHz)
$p_Y(y)$	0.4	0.2	0.4

Expectation

Definition

The expected value of a function $h(x,y)$ is

$$E[h(X,Y)] = \sum_{x,y} h(x,y) p_{X,Y}(x,y).$$

Example (cont.)

What is $E[|X-Y|]$ (the average speed difference)?

Here, we have the situation $E[|X-Y|] = E[h(X,Y)]$, with $h(X,Y) = |X-Y|$. Thus, we have

$$
\begin{aligned}
E[|X-Y|] &= \textstyle\sum_{x,y} |x-y| p_{X,Y}(x,y) = \\
&= |400-400| \cdot 0.1 + |400-450| \cdot 0.1 + |400-500| \cdot 0.2 \\
&+ |450-400| \cdot 0.1 + |450-450| \cdot 0.0 + |450-500| \cdot 0.1 \\
&+ |500-400| \cdot 0.2 + |500-450| \cdot 0.1 + |500-500| \cdot 0.1 \\
&= 0+5+20+5+0+5+20+5+0 = 60.
\end{aligned}
$$

Covariance

Definition

The covariance between two random variables X and Y is

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

This definition looks very much like the definition for the variance of a single random variable. In fact, if we set $Y = X$ in the above definition, then $Cov(X,X) = Var(X)$.

Correlation

Definition

The correlation between two variables X and Y is

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

Properties:

- ρ is between -1 and 1
- if $\rho = 1$ or -1, Y is a linear function of X
  $\rho = 1 \quad \to Y = aX + b$ with $a > 0$,
  $\rho = -1 \quad \to Y = aX + b$ with $a < 0$,
- ρ is a measure of linear association between X and Y
- ρ near ± 1 indicates a strong linear relationship, ρ near 0 indicates lack of linear association.

Example (cont.)

What is $\rho(X,Y)$ in our box with five chips?

Use marginal pmfs to compute:

- $E[X] = E[Y] = 450$
- $Var[X] = Var[Y] = 2000$

The covariance between X and Y is:

$$\begin{aligned} Cov(X,Y) &= \textstyle\sum_{x,y}(x - E[X])(y - E[Y])p_{X,Y}(x,y) = \\ &= (400 - 450)(400 - 450) \cdot 0.1 + (450 - 450)(400 - 450) \cdot 0.1 + \\ &\quad \cdots + (500 - 450)(500 - 450) \cdot 0.1 \\ &= 250 + 0 - 500 + 0 + 0 + 0 - 500 + 250 + 0 = -500. \end{aligned}$$

The correlation there is

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{-500}{2000} = -0.25,$$

and thus there is a weak negative (linear) association.

Example (cont.)

Definition

Two discrete random variables are independent if

$$P(X=x, Y=y) = P(X=x)P(Y=y).$$

Are X and Y independent?

Intuition: No, since if we know X then it will change what we think about Y.

Definition: independence if $p_{x,y}(x,y) = p_X(x)p_Y(y)$ for all x and y.

Since

$$p_{X,Y}(450, 450) = 0 \neq 0.2 \cdot 0.2 = p_X(450) \cdot p_Y(450)$$

they are **not** independent.

Continuous random variables

All the properties have continuous analogs.

Definition

Suppose X and Y are two continuous random variables with joint probability density function $p_{X,Y}(x, y)$. Then the marginal probability density functions are

$$\begin{aligned} p_X(x) &= \int p_{X,Y}(x, y)dy \\ p_Y(y) &= \int p_{X,Y}(x, y)dx. \end{aligned}$$

Two continuous random variables are independent if

$$p_{X,Y}(x, y) = p_X(x)p_Y(y).$$

The expected value is

$$E[h(X, Y)] = \int \int h(x, y)p_{X,Y}(x, y)dxdy.$$

Properties of variances and covariances

For any random variables X, Y, W and Z,

$$\begin{aligned}
Var(aX+bY+c) &= a^2Var(X)+b^2Var(Y)+2abCov(X,Y) \\
Cov(aX+bY,cZ+dW) &= acCov(X,Z)+adCov(X,W) \\
&\quad +bcCov(Y,Z)+bdCov(Y,W) \\
Cov(X,Y) &= Cov(Y,X) \\
\rho(X,Y) &= \rho(Y,X)
\end{aligned}$$

If X and Y are independent, then

$$\begin{aligned}
Cov(X,Y) &= 0 \\
Var(aX+bY+c) &= a^2Var(X)+b^2Var(Y).
\end{aligned}$$