Jointly Distributed Random Variables

CE 311S

What if there is more than one random variable we are interested in?

How should you invest the extra money from your summer internship?

To simplify matters, imagine there are two mutual funds you are thinking about investing in:

Fund A: Invests in cryptocurrency

Fund B: Invests in municipal bonds

Neither of these funds has a guaranteed rate of return, so we can use probability distributions to describe them.

Based on your investing experience, you believe that Fund A's annual rate of return will be 75% with probability 0.2, 20% with probability 0.5, and -50% with probability 0.3.

Fund B has an annual rate of return of 10% with probability 0.6, and -5% with probability 0.4.

However, investments are not independent of one another: when the economy is strong, most assets will increase in value; in a recession, most assets will decrease in value.

We can describe this information in a table showing the probability of seeing *each combination* of rates of return.

B	A			
	+75%	+20%	−50%	Sum
+10%	0.10	0.45	0.05	0.6
−5%	0.10	0.05	0.25	0.4
Sum	0.20	0.50	0.30	1

Each entry in the table shows the probability of a particular combination of rates of return. This is called the **joint probability mass function** or **joint distribution** of A and B.

B	A			
	$+75\%$	$+20\%$	-50%	Sum
$+10\%$	0.10	0.45	0.05	0.6
-5%	0.10	0.05	0.25	0.4
Sum	0.20	0.50	0.30	1

Some things to notice about the table:

- Each value is nonnegative, and all values in the table add up to 1.
- The sum of all values in the first row gives the probability that $B = +10\%$ when we aren't looking at A.
- The sum of the values in each column give the probability mass function for A when we aren't looking at B.

In general, if X and Y are any two discrete variables, the **joint probability mass function** $P_{XY}(x, y)$ is the probability of seeing both $X = x$ and $Y = y$.

To be a valid joint PMF, $P_{XY}(x, y) \geq 0$ for all x and y, and $\sum_x \sum_y P_{XY}(x, y) = 1$.

The **marginal PMF of** X gives us the distribution of X when we aren't concerned with Y:

$$P_X(x) = \sum_y P_{XY}(x, y)$$

Likewise, the marginal PMF of Y is $P_Y(y) = \sum_x P_{XY}(x, y)$

B	A			
	+75%	+20%	−50%	Sum
+10%	0.10	0.45	0.05	0.6
−5%	0.10	0.05	0.25	0.4
p_A	**0.20**	**0.50**	**0.30**	1

The marginal PMF of A in this table is just the sums of each column.

B	A			
	$+75\%$	$+20\%$	-50%	p_B
$+10\%$	0.10	0.45	0.05	**0.6**
-5%	0.10	0.05	0.25	**0.4**
p_A	0.20	0.50	0.30	1

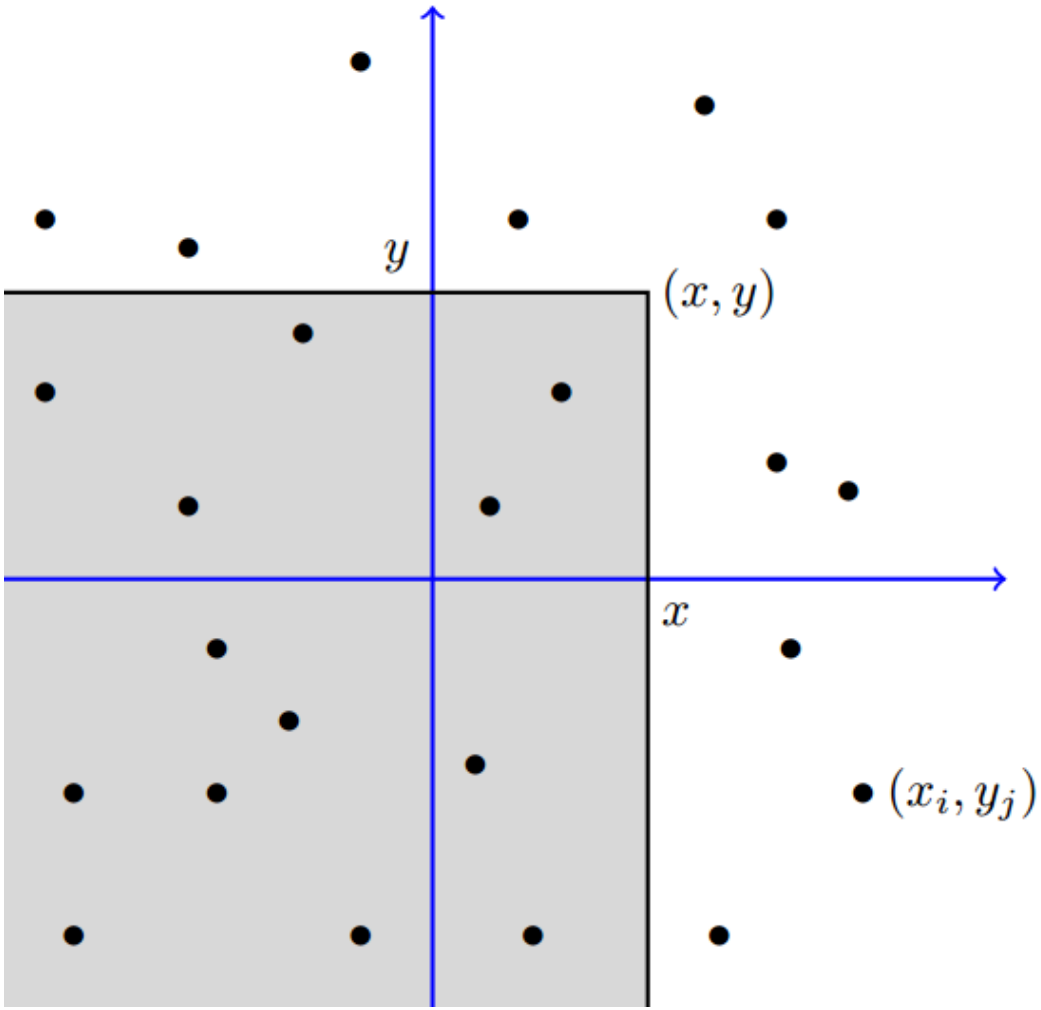
The marginal PMF of B is just the sums of each row.

The joint CDF is written

$$F_{XY}(x, y) = P(X \leq x \cap Y \leq y)$$

and is the sum of the P_{XY} values which are *both* less than or equal to x, and less than or equal to y.

To use the joint CDF to solve problems, it is helpful to draw a picture to see what areas need to be added and subtracted.



(Figure from Pishro-Nik)

Two discrete random variables X and Y are **independent** if $P_{XY}(x, y) = P_X(x)P_Y(y)$ for *every* possible value of x and y. If this is not true (even for one value of x and y), they are **dependent**.

B	A			
	$+75\%$	$+20\%$	-50%	P_B
$+10\%$	0.10	0.45	0.05	0.6
-5%	0.10	0.05	0.25	0.4
P_A	0.20	0.50	0.30	1

Are funds A and B independent?

Let R be the number of heads on three coin flips, and S be the number of heads on the *next* two coin flips.

S	R				
	0	1	2	3	P_S
0	1/32	3/32	3/32	1/32	1/4
1	1/16	3/16	3/16	1/16	1/2
2	1/32	3/32	3/32	1/32	1/4
P_R	1/8	3/8	3/8	1/8	1

Are R and S independent?

How did I get the values for $P_{RS}(r, s)$?

The expected value of any function $h(X, Y)$ is $\sum_x \sum_y h(x, y) p_{XY}(x, y)$

B	A			
	$+75\%$	$+20\%$	-50%	p_B
$+10\%$	0.10	0.45	0.05	0.6
-5%	0.10	0.05	0.25	0.4
p_A	0.20	0.50	0.30	1

Let's say I invest my money equally in the two funds. What is the expected average rate of return $(A + B)/2$?

To calculate the sum, you can create two tables, one with values of $(A + B)/2$, and the other with probabilities.

$(A + B)/2$**:**

B	A		
	$+75\%$	$+20\%$	-50%
$+10\%$	$+42.5\%$	$+15\%$	-20%
-5%	$+35\%$	$+7.5\%$	-27.5%

Probabilities:

B	A		
	$+75\%$	$+20\%$	-50%
$+10\%$	0.10	0.45	0.05
-5%	0.10	0.05	0.25

Multiply corresponding values and add:
$(42.5 \times 0.10 + 15 \times 0.45 + \cdots + (-27.5) \times 0.25) = \mathbf{7}$.

What are $E[A]$ and $E[B]$? **Probabilities:**

B	A		
	$+75\%$	$+20\%$	-50%
$+10\%$	0.10	0.45	0.05
-5%	0.10	0.05	0.25

A:

B	A		
	$+75\%$	$+20\%$	-50%
$+10\%$	$+75\%$	$+20\%$	-50%
-5%	$+75\%$	$+20\%$	-50%

Probabilities:

B	A		
	$+75\%$	$+20\%$	-50%
$+10\%$	0.10	0.45	0.05
-5%	0.10	0.05	0.25

Multiply corresponding values and add:
$(75 \times 0.10 + 20 \times 0.45 + \cdots + (-50) \times 0.25) = \mathbf{10}$.
(You would get the same answer by calculating $E[A]$ the usual way, from the *marginal distributions*.)

B:

B	A		
	+75%	+20%	−50%
+10%	+10%	+10%	+10%
−5%	−5%	−5%	−5%

Probabilities:

B	A		
	+75%	+20%	−50%
+10%	0.10	0.45	0.05
−5%	0.10	0.05	0.25

Multiply corresponding values and add:
$(10 \times 0.10 + 10 \times 0.45 + \cdots + (-5) \times 0.25) = \mathbf{4}$.
(You would get the same answer by calculating $E[B]$ the usual way, from the *marginal distributions*.)

MULTIPLE CONTINUOUS RANDOM VARIABLES

All of the same concepts can be applied for joint continuous random variables as well.

The **joint density function** $f_{XY}(x, y)$ is valid if $f_{XY}(x, y) \geq 0$ for all x and y, and if $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y)\ dy\ dx = 1$.

The **marginal density functions** are $f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\ dy$ and $f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\ dx$

X and Y are independent if $f_{XY}(x, y) = f_X(x) f_Y(y)$ for all x and y

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

Continuous joint random variables are similar, but let's go through some examples. (Key ideas: replace table of "masses" with a "density" function of multiple variables; replace sums with integrals)

You have a new laptop. Let X represent the time (in years) before the hard drive fails, and Y the time in years before your keyboard breaks. Let the joint PDF of X and Y be

$$f_{XY}(x, y) = ke^{-3x-2y}$$

for $x > 0$ and $y > 0$, and 0 otherwise. What is k?

So $f_{XY}(x, y) = 6e^{-3x-2y}$ for $x > 0$, $y > 0$.
What is the marginal distribution of X? (This gives us the pdf for failure life of the hard drive, irrespective of the failure life of the keyboard.)

$f_{XY}(x, y) = 6e^{-3x-2y}$ for $x > 0$, $y > 0$.
What is the marginal distribution of Y? (This gives us the pdf for failure life of the keyboard, irrespective of the failure life of the hard drive.)

X and Y are independent if $f_{XY}(x, y) = f_X(x)f_Y(y)$ Are X and Y independent?

What is the expected lifetime of the keyboard?

Assume that I throw the laptop away once either the keyboard or hard drive break. What is the expected lifetime of the laptop?

COVARIANCE AND CORRELATION

In the discrete example, we already saw that funds A and B are not independent. It would be useful to have a measure of *how* dependent they are, though.

Define the **covariance** of two random variables X and Y to be $\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$.

Why does this tell us how dependent X and Y are?

There is a shortcut formula for covariance:
$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$.

Recall that $E[A] = 10$ and $E[B] = 4$. $(A - E[A])(B - E[B])$:

B	A		
	$+75\%$	$+20\%$	-50%
$+10\%$	390	60	-360
-5%	-585	-90	540

Probabilities:

B	A		
	$+75\%$	$+20\%$	-50%
$+10\%$	0.10	0.45	0.05
-5%	0.10	0.05	0.25

Multiply corresponding values and add:
$(390 \times 0.10 + 60 \times 0.45 + \cdots + 540 \times 0.25) = \mathbf{120}$.

Positive covariance means that when A is high, B tends to be high; and when A is low, B tends to be low as well.

We can also use the shortcut formula $E[AB] - E[A]E[B]$ to save some tedious computations: AB**:**

B	A		
	$+75\%$	$+20\%$	-50%
$+10\%$	750	200	-500
-5%	-375	-100	250

Probabilities:

B	A		
	$+75\%$	$+20\%$	-50%
$+10\%$	0.10	0.45	0.05
-5%	0.10	0.05	0.25

$E[AB] = 160$, so $\text{Cov}(A, B) = 160 - 4 \times 10 = 120$

The trouble with covariance by itself is that it is hard to interpret, apart from the sign:

- If covariance is positive, when X is above average, Y usually is too; and when X is below average, Y usually is too.
- If covariance is negative, when X is above average, Y is usually below average, and vice versa.
- If X and Y are independent, their covariance is zero. (The converse is **not** true).

The magnitude of the covariance does not tell very much, though. (It depends on the units of X and Y)

To make the magnitude useful, we make covariance unitless by dividing by the standard deviations of X and Y. This gives the **correlation coefficient**:

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

The correlation coefficient is always between -1 and $+1$, and quantifies the strength of the **linear** relationship between X and Y.

- If $\rho_{XY} = 1$, then $Y = aX + b$ for some $a > 0$.
- If $\rho_{XY} = -1$, then $Y = aX + b$ for some $a < 0$.
- If $\rho_{XY} = 0$, there is no linear relationship between X and Y. (This does **not** mean they are independent.)
- If $\rho_{XY} \in (0, 1)$, there is roughly a linear relationship with positive slope.
- If $\rho_{XY} \in (-1, 0)$, there is roughly a linear relationship with negative slope.

In the example with the mutual funds, $\sigma_A = 44.4$ and $\sigma_B = 7.35$, so $\rho_{AB} = 120/(44.4 \times 7.35) = 0.367$

Properties of covariance

1. $\text{Cov}(X, X) = \text{Var}[X]$ (covariance of a random variable with itself is just its variance)
2. X, Y independent $\Rightarrow \text{Cov}(X, Y) = 0$ (but the reverse is **not** always true)
3. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$ (order doesn't matter for covariance)
4. $\text{Cov}(aX, Y) = a\text{Cov}(X, Y)$ (constants can be factored out)
5. $\text{Cov}(X + c, Y) = \text{Cov}(X, Y)$ (adding a constant does not change covariance)
6. $\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$ (distributive property)

(You should be able to prove these formulas from the definitions given above.)

By combining these properties, we can obtain the general formula

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

For instnace, $\text{Cov}(X_1 + 2X_2, 3Y_1 + 4Y_2) =$
$3\text{Cov}(X_1, Y_1) + 4\text{Cov}(X_1, Y_2) + 6\text{Cov}(X_2, Y_1) + 8\text{Cov}(X_2, Y_2)$

Let X and Y be independent standard normal random variables. What is $\text{Cov}(1 + X + XY^2, 1 + X)$?

1. Added constants can be removed: $\text{Cov}(X + XY^2, X)$
2. Distributive property: $\text{Cov}(X, X) + \text{Cov}(XY^2, X)$
3. Covariance with itself is variance: $\text{Var}(X) + \text{Cov}(XY^2, X)$
4. Standard normal has variance 1: $1 + \text{Cov}(XY^2, X)$
5. Shortcut formula: $1 + E[X^2Y^2] - E[XY^2]E[X]$
6. Independence: $1 + E[X^2]E[Y^2] - E[XY^2]E[X]$
7. Variance shortcut formula:
   $1 + (\text{Var}(X) + E[X]^2)(\text{Var}(Y) + E[Y]^2) - E[XY^2]E[X]$
8. Standard normal: $1 + (1 + 0)(1 + 0) - E[XY^2](0) = 2$

An important special case is finding the variance of a sum:

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y)$$

What does this mean if X and Y are independent? Positively correlated? Negatively correlated?