Chapter 4: Joint probability distributions

Mariona Segú

Assistant Professor
Thema, CY Cergy Paris Université

Fall 2023

Contents

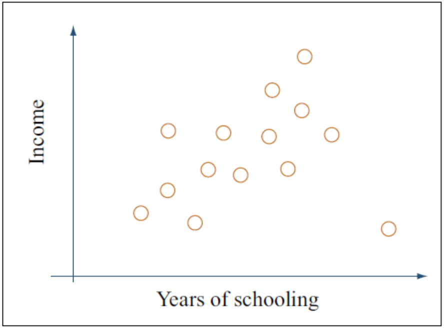
Contents

1. Joint Distributions of 2 Random Variables X, Y

Often, we are interested not only in a single random variable, but about the relationship between two or more variables, e.g. whether the outcome of some process affects the outcome of another.
Examples:

- IQs of identical twins -i.e. X would be one kid's IQ, and Y that of her/his sibling
- Educational attainment X and income Y: we can plot both variables together.

1. Joint Distributions of 2 Random Variables X, Y



In the graph it looks like there is in fact a non-trivial relationship between the variables.

Contents

2. Joint Distributions: Discrete Random Variables

In the **discrete** case, the joint PDF is given by

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

As usual,

- $p(x, y) \geq 0 \qquad$ for all x, y
- For all values $(x, y) \in \mathbb{R}^2$.

$$\sum_{i=1}^{n} p(x_i, y_i) = 1$$

2. Joint Distributions: Discrete Random Variables

Die tossing

Imagine we toss two die, X is the outcome of dice 1 and Y the outcome of the second. What is $P(2 \leq X \leq 3, 1 \leq Y \leq 2)$?

2. Joint Distributions: Discrete Random Variables

Die tossing

Imagine we toss two die, X is the outcome of dice 1 and Y the outcome of the second. What is $P(2 \leq X \leq 3, 1 \leq Y \leq 2)$?

$$P(2 \leq X \leq 3, 1 \leq Y \leq 2) = p(2,1) + p(2,2) + p(3,1) + p(3,2)$$

Since the probability of each individual pair is p(x,y)=1/36 and all pairs have the same probability...

$$P(2 \leq X \leq 3, 1 \leq Y \leq 2) = 4/36 = 1/9$$

2. Joint Distributions: Discrete Random Variables

Queuing in the supermarket

In a supermarket, let X be the number of people in the regular checkout line, and Y the number of people in the express line. Then the joint PDF of X and Y could look like this:

f_{XY}		$Y=0$	$Y=1$	$Y=2$	$Y=3$	$Y=4$	Total
X	0	0.1	0.05	0.05	0	0	0.2
	1	0.05	0.2	0.2	0.05	0	0.5
	2	0	0	0.1	0.1	0.05	0.25
	3	0	0	0	0	0.05	0.05
		0.15	0.25	0.35	0.15	0.1	1

2. Joint Distributions: Discrete Random Variables

Queuing in the supermarket

This is a **contingency table**:

- Cell probabilities from the joint PDF
- Marginal probabilities on the sides
- The sum of all probabilities is equal to 1

Note that: when the number of individuals at the regular checkout is high, then the number of persons in the express line also tends to be high.

2. Joint Distributions: Discrete Random Variables

Queuing in the supermarket

We can also calculate probabilities for different events based on the PDF as given in the table:

$$P(X = 2)$$

$$P(X \geq 2, Y \geq 2)$$

$$P(|X - Y| \leq 1)$$

2. Joint Distributions: Discrete Random Variables

Queuing in the supermarket

We can also calculate probabilities for different events based on the PDF as given in the table:

$$P(X = 2) = 0 + 0 + 0.1 + 0.1 + 0.05 = 0.25$$

$$P(X \geq 2, Y \geq 2) = \sum_{x=2}^{3} \sum_{y=2}^{4} f(x, y)$$
$$= 0.1 + 0.1 + 0.05 + 0 + 0 + 0.05 = 0.3$$

$$P(|X - Y| \leq 1) = P(X = Y) + P(|X - Y| = 1)$$
$$= 0.1 + 0.2 + 0.1 + 0 + 0.05 + 0.05 + 0.2 + 0 + 0.1 + 0 + 0.05 = 0.85$$

2. Joint Distributions: Discrete Random Variables

The joint CDF of two discrete random variables is:

$$F(a,b) = P(X \leq a, Y \leq b) = \sum_{x=-\infty}^{a} \sum_{y=-\infty}^{b} p(x,y)$$

Die tossing

$$F(2,3) = P(X \leq 2, Y \leq 3)$$

$$F(2,3) = p(1,1) + p(1,2) + p(1,3) + p(2,1) + p(2,2) + p(2,3)$$

Hence $F(2,3) = 6/36 = 1/6$

Contents

3. Joint Distributions: Continuous Random Variables

If X and Y are continuous random variables, we note the joint PDF of (X, Y) as $f(x, y)$. Their CDF is then

$$F(a, b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y) dx dy$$

3. Joint Distributions: Continuous Random Variables

As for the single-variable case, the PDF must satisfy:

- Any single point has probability zero
- $$f(x,y) \geq 0 \qquad \text{for each}(x,y) \in \mathbb{R}^2$$
- $$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$$

3. Joint Distributions: Continuous Random Variables

The CDF must satisfy:

- $F(-\infty,-\infty) = F(-\infty,y) = F(x,-\infty) = 0$
- $F(\infty,\infty) = 1$
- If $a_2 \geq a_1$ and $b_2 \geq b_1$ then

$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) \geq 0$$

which is the same as

$$F(a_2,b_2) - F(a_2,b_1) - F(a_1,b_2) + F(a_1,b_1) \geq 0$$

We have excluded the events where $X \leq a_1$ and $Y \leq b_1$ twice. So, to adjust for this over subtraction, we need to add back $F(a_1,b_1)$.

3. Joint Distributions: Continuous Random Variables

UFOs

A UFO appears at a random location over Wyoming, a rectangle of 276 times 375 miles. The position of the UFO is uniformly distributed over the entire state, and can be expressed as a random longitude X (from -111 to -104 degrees) and latitude Y (between 41 and 45 degrees). The joint density of the coordinates is given by

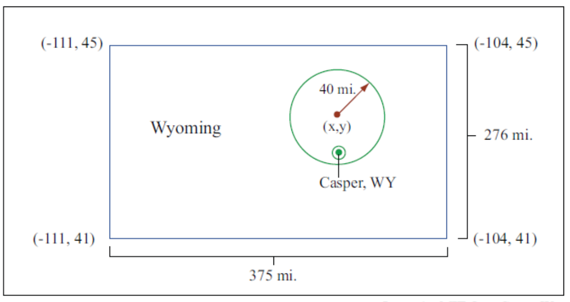
$$f(x,y) = \begin{cases} \frac{1}{28} & \text{if } -111 \leq x \leq -104 \text{ and } 41 \leq y \leq 45 \\ 0 & \text{otherwise} \end{cases}$$

Since there are $7 \times 4 = 28$ possible combinations of X and Y

3. Joint Distributions: Continuous Random Variables

UFOs

If the UFO can be seen from a distance of up to 40 miles, what is the probability that it can be seen from Casper,WY (in the middle of the state)? The set of locations for which the UFO can be seen from Casper are a 40-mile radius circle around Casper. Since the density is uniform we can use geometry to compute the density function:



3. Joint Distributions: Continuous Random Variables

UFOs

We can calculate the probability as

$$\text{P}(\text{"}{<}40\text{m Casper"}) = \frac{\text{Area}(\text{"}{<}40\text{m Casper"})}{\text{Area}(\text{"All of Wyoming"})} = \frac{40^2\pi}{375 \cdot 276} = 0.049$$

Notice that for the uniform distribution, there is no need to perform complicated integration, you can treat everything as a purely geometric problem.

Contents

Marginal Probability

If we have a joint distribution, we may want to recover the distribution of one variable X.

Definition If X and Y are discrete random variables with joint PF f, then, the **marginal probability** is

$$p_X(x) = \sum_{\text{all } y} p(x, y) \qquad p_Y(y) = \sum_{\text{all } x} p(x, y)$$

Note that $p(.)$ has the sub-index X to indicate "marginal".

Marginal Probability - Discrete Case

Die toss

Take the example of two die X and Y. If we want to recover the probability of $X = 1$, we need to count all the X, Y combinations that have $X = 1$. This is, we need to count over all possible values of Y:

$$P(X = 1) = p(1,1) + p(1,2) + ... + p(1,6) = 1/36 \times 6 = 1/6$$

Expressed in summation notation:

$$p_X(X) = \sum_{y=1}^{6} p(x, y)$$

Marginal Probability - Discrete Case

Happy marriage?

In a survey, individuals rate their marriage from 1 (unhappy) to 3 (happy), and report the number of years married. Look at the joint distribution of "marriage quality", X, and duration Y, the "cell" probabilities and the marginal PDFs:

		Y			
	f_{XY}	1	8	12	f_X
X	1	4.66%	11.48%	12.98%	29.12%
	2	5.16%	14.81%	12.31%	32.28%
	3	13.48%	16.47%	8.65%	38.60%
	f_Y	23.30%	42.76%	33.94%	100.00%

Note that the joint distribution is concentrated along the bottom left/top right diagonal. Back (Indep)

Conditional Probability

Let's look at *conditional* distributions:
According to the multiplicative law (CH1), we have:

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$$

Let's apply this to the intersection of event $X = x$ and $Y = y$:

$$p(x, y) = p_X(x)p(y|x) = p_Y(y)p(x|y)$$

where $p(x|y)$ is the probability that $X = x$ given Y takes the value y

Conditional Probability

Definition The **conditional PDF** of Y given X is

$$p(y|x) = \frac{P(Y = y|X = x)}{P(X = x)} = \frac{p(x,y)}{p_X(x)}$$

Conditional Probability - Discrete Case

Happy marriage?

Let's now look at the number of extra-marital affairs during the last year, Z, and self-reported marriage quality, X. The joint PDF is given by

		Z			
	f_{XZ}	0	1	2	f_X
X	1	17.80%	4.49%	6.82%	29.12%
	2	24.29%	3.83%	4.16%	32.28%
	3	32.95%	3.33%	2.33%	38.60%
	f_Z	75.04%	11.65%	13.31%	100.00%

Conditional Probability - Discrete Case

Happy marriage?

It is more instructive to look at the PDF of the number of affairs Z conditional on the rating of marriage quality. Conditional on the low rating, $X = 1$, we have

$$f_{Z|X}(0|1) = \frac{f_{ZX}(0,1)}{f_X(1)} = \frac{17.8}{29.12} = 61.13$$

	$f_{Z\|X}$	Z = 0	1	2
X	1	61.13%		
	2			
	3			

Conditional Probability - Discrete Case

Happy marriage?

	$f_{Z\|X}$	Z = 0	Z = 1	Z = 2
X	1	61.13%	15.42%	23.42%
	2	75.25%	11.86%	12.88%
	3	85.36%	8.63%	6.04%

we can see that for lower values of marriage quality X, the conditional PDF puts higher probability mass on higher numbers of affairs.

Conditional Probability - Discrete Case

Happy marriage?

Does this mean that dissatisfaction with marriage causes extra-marital affairs? Certainly not: let's look at the reverse exercise $(X|Z)$

$$f_{X|Z}(1|0) = \frac{f_{XZ}(1,0)}{f_Z(0)} = \frac{17.8}{75.04} = 23.72$$

	$f_{X\|Z}$	Z = 0	1	2
	1	23.72%	38.54%	51.24%
X	2	32.37%	32.88%	31.25%
	3	43.91%	28.58%	17.51%

Conditional Probability - Discrete Case

Happy marriage?

So we could as well read the numbers as extra-marital affairs ruining the relationship.
This is often referred to as **"reverse causality"** : even though we may believe that A causes B, B may at the same time cause A.

Marginal Probability —Continuous Case

Definition. If X and Y are continuous random variables with joint PDF $f(x,y)$ defined over a region $\mathcal{R}$, the **marginal densities** are obtained by integrating over the range of the other variable:

$$f_X(x) = \int_{\text{all } y} f(x,y)\,dy, \qquad f_Y(y) = \int_{\text{all } x} f(x,y)\,dx.$$

$$f_X(x) = \int_a^b f(x,y)\,dy, \qquad f_Y(y) = \int_a^b f(x,y)\,dx.$$

Conditional Probability - Continuous Case

For two continuous random variables X and Y, their conditional PDF is:

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} \qquad \text{if } f_Y(y) > 0$$

Similarly,

$$f(y|x) = \frac{f(x,y)}{f_X(x)} \qquad \text{if } f_X(x) > 0$$

4. Marginal and Conditional Probability

A water dispenser

A water machine has a random amount Y in supply at the beginning of a given day and dispenses a random amount X during the day (with measurements in liters). It is not resupplied during the day, and hence $Y \geq X$. It has been observed that Y and X have a joint density given by

$$f(x, y) = 1/2 \qquad \text{if } 0 \leq x \leq y; 0 \leq y \leq 2,$$

Find the two marginal probabilities

Solution: Marginal densities

A water dispenser

$$f(x,y) = \tfrac{1}{2} \quad \text{for } 0 \leq x \leq y \leq 2, \qquad 0 \text{ otherwise.}$$

Marginal of X:

$$f_X(x) = \int_{\text{all } y} f(x,y)\,dy = \int_{y=x}^{2} \tfrac{1}{2}\,dy = \tfrac{1}{2}(2-x), \quad 0 \leq x \leq 2.$$

Marginal of Y:

$$f_Y(y) = \int_{\text{all } x} f(x,y)\,dx = \int_{x=0}^{y} \tfrac{1}{2}\,dx = \tfrac{y}{2}, \quad 0 \leq y \leq 2.$$

Check: $\int_0^2 f_X(x)\,dx = 1$ and $\int_0^2 f_Y(y)\,dy = 1$.

4. Marginal and Conditional Probability

A water dispenser

A water machine has a random amount Y in supply at the beginning of a given day and dispenses a random amount X during the day (with measurements in liters). It is not resupplied during the day, and hence $Y \geq X$. It has been observed that Y and X have a joint density given by

$$f(x, y) = 1/2 \qquad \text{if } 0 \leq x \leq y \leq 2,$$

Find the conditional density of X given $Y = y$. Evaluate the probability that less than $1/2$ liter will be used, given that the machine contains 1.5 liter at the start of the day.

4. Marginal and Conditional Probability

A water dispenser

We need to find

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} \qquad \text{if } 0 \leq x \leq y \leq 2$$

We know $f(x,y)$. We know $f_Y(y)$

$$f_Y(y) = (1/2)y$$

Hence

$$f(x|y) = \frac{1/2}{y/2} = \frac{1}{y} \qquad \text{if } 0 \leq x \leq y \leq 2$$

4. Marginal and Conditional Probability

A water dispenser

Now, we can find $P(X \leq 1/2|Y = 1.5)$.

$$P(X \leq 1/2|Y = 1.5) = \int_a^{1/2} f(x|y = 1.5)dx = \int_0^{1/2} (\frac{1}{1.5})dx = \frac{1/5}{1.5} = \frac{1}{3}$$

Contents

5. Independence

Two events A and B are independent if $P(AB) = P(A)P(B)$.
Let's define a similar notion for random variables.
Definition We say that the random variables X and Y are **independent** if for any regions $A, B \subset R$,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

Very strict requirement!: it means that all pairs of events are mutually independent. This definition is difficult to check...

5. Independence

However if X and Y are independent, it follows that

$$F(x,y) = P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) = F(x)F(y)$$

From which we can derive:

Proposition X and Y are independent if and only if their joint and marginal PDFs satisfy

- Discrete case: $p(x,y) = p_X(x)p_Y(y)$
- Continuous case: $f(x,y) = f_X(x)f_Y(y)$

In words: learning the value of Y doesn't change any of the probabilities associated with X

5. Independence

Die Tossing

Let's go back to the 2 die tossing. Show that X and Y are independent.

Each of the 36 combinations has a probability $1/36$. For example $p(1,2) = 1/36$. At the same time; $p_X(1) = 1/6$ and $p_Y(2) = 1/6$. Hence,

$$1/36 = p(1,2) = p_X(1)p_Y(2) = 1/6 \times 1/6 = 1/36$$

They are then, independent

5. Independence

Happy marriage? 2

In the previous example, we calculated the **marginal PDFs** of reported "marriage quality", X, and years married, Y as

$$f_X(1) = 29.12, \qquad f_X(2) = 32.28, \qquad f_X(3) = 38.60$$

and

$$f_Y(1) = 23.30, \qquad f_Y(8) = 42.76, \qquad f_Y(12) = 33.94$$

Let's check if they are independent:

$$\text{Is } f(3,1) = f_X(3)f_Y(1)?$$

▸ Table

5. Independence

Happy marriage? 2

$$13.48 = f(3,1) \neq f_X(3)f_Y(1) = 38.6 \cdot 23.3 = 8.99$$

They are not independent! Let's check the whole table

	$\tilde{f}_{XY}$	Y: 1	8	12	f_X
X	1	6.78%	12.45%	9.88%	29.12%
	2	7.52%	14.81%	10.96%	32.28%
	3	8.99%	16.50%	13.10%	38.60%
	f_Y	23.30%	42.76%	33.94%	100.00%

5. Independence

A Continuous example

$$f(x,y) = 4xy \qquad \text{if } 0 \leq x \leq 1; 0 \leq y \leq 1$$

Show that X and Y are independent.

5. Independence

A Continuous example

$$f(x, y) = 4xy \qquad \text{if } 0 \le x \le 1; 0 \le y \le 1$$

Show that X and Y are independent.

$$f_X(x) = \int_a^b f(x, y)dy = \int_0^1 (4xy)dy = 2xy^2\Big]_0^1 = 2x$$

Similarly, $f_Y(y) = 2y$. Hence,

$$4xy = f(x, y) = f_X(x)f_Y(y) = 2x \times 2y = 4xy$$

So X and Y are independent

5. Independence

Remark If the limits of the integration are constants (and not variables), the independence condition can be restated as follows: Whenever we can factor the joint PDF into

$$f(x, y) = g(x)h(y)$$

where $g(\cdot)$ depends only on x and $h(\cdot)$ depends only on y, then X and Y are independent.
In particular, we don't have to calculate the marginal densities explicitly.

5. Independence

Example

Say, we have a joint PDF

$$f(x, y) = ce^{-(x+2y)} \qquad \text{if } x \geq 0, y \geq 0$$

Then we can choose e.g. $g(x) = ce^{-x}$ and $h(y) = e^{-2y}$, and even though these aren't proper densities, this is enough to show that X and Y are independent.

5. Independence

Example 2

Suppose we have the joint PDF

$$f(x,y) = cx^2y \qquad \text{if } x \leq y \leq 1$$

Can X and Y be independent?
Even if the PDF factors into functions of x and y, we can also see that the support of X depends on the value of Y, and therefore, X and Y can't be independent.

Contents

6. Expected Value of a Function of a Random Variables

We can apply the same definition of Expected Value for the multivariate case:
If $g(X, Y)$ is a function of RV X and Y with PDF $p(X, Y)$, the expected value of $g(X, Y)$ is

$$E[g(X,Y)] = \sum_X \sum_Y g(x,y)p(x,y)$$

in the discrete case, and

$$E[g(X,Y)] = \int_X \int_Y g(x,y)f(x,y)dxdy$$

in the continuous case

6. Expected Value of a Function of a Random Variables

In general, we will be interested in computing

$$E[XY] = \sum_X \sum_Y xyp(x,y)$$

in the discrete case, and

$$E[XY] = \int_X \int_Y xyf(x,y)dxdy$$

in the continuous case. For single expectation, we use:

$$E[X] = \int_X xf_X(x)dx \qquad \text{and } E[X] = \sum_X xp(x)$$

6. Expected Value of a Function of a Random Variables

Example

		Y		
		0	1	2
	0	1/9	2/9	1/9
X	1	2/9	2/9	0
	2	1/9	0	0

Find $E(Y)$

6. Expected Value of a Function of a Random Variables

Example

Find $E(Y)$:

		Y			$p_X(x)$
		0	1	2	
	0	1/9	2/9	1/9	
X	1	2/9	2/9	0	
	2	1/9	0	0	
	$p_Y(y)$	4/9	4/9	1/9	

$$E(Y) = \sum_{0}^{2} y_i p_y(y) = 0 \times p_Y(0) + 1 \times p_Y(1) + 2 \times p_Y(2)$$

$$= 4/9 + 2 \times 1/9 = 6/9 = 2/3$$

6. Expected Value of a Function of a Random Variables

Example

Find $E(XY)$:

$$E(XY) = \sum_{i=0}^{2}\sum_{j=0}^{2} x \times y \times P(X = x, Y = y)$$

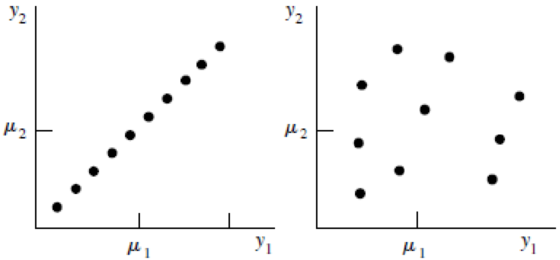
Now, filling in the values from the table:

$$\begin{aligned} E(XY) = 0 \times 0 \times \frac{1}{9} + 0 \times 1 \times \frac{2}{9} + 0 \times 2 \times \frac{1}{9} + 1 \times 0 \times \frac{2}{9} \\ +1 \times 1 \times \frac{2}{9} + 1 \times 2 \times 0 + 2 \times 0 \times \frac{1}{9} + 2 \times 1 \times 0 \\ +2 \times 2 \times 0 = \frac{2}{9} \end{aligned}$$

Contents

7. The Covariance

We use two measurements to measure the dependence between two variables: the covariance and the correlation coefficient.

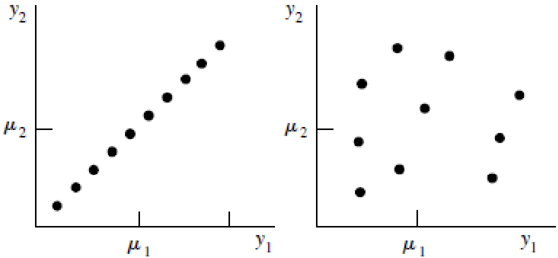


(a) (b)

Let's compute $(y_2 - \mu_2)$ and $(y_1 - \mu_1)$ for each point

7. The Covariance

In Figure (a), the points where $(y_2 - \mu_2) < 0$ will also have $(y_1 - \mu_1) < 0$. Which means that the product $(y_2 - \mu_2)(y_1 - \mu_1)$ will always be positive.



(a) (b)

In (b), the product will be sometimes positive, sometimes negative and will have an average close to zero.

7. The Covariance

The previous example shows that the sign of the average of $(y_2 - \mu_2)(y_1 - \mu_1)$ is informative:

Definition The **covariance** between X and Y is:

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

where $\mu_X = E(X)$ and $\mu_Y = E(Y)$.

The larger the absolute value of $Cov(X, Y)$, the greater the linear dependence between X and Y. Positive values indicate that X increases as Y increases; and the opposite is true.

7. The Covariance

Yet, the covariance depends on the scale of measurements. It is hard to use it as an absolute measure. This can be solve by standarizing the value.

Definition The **coefficient of correlation** between X and Y is:

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are the standard deviations.

The Correlation has a range of $-1 \leq \rho \leq 1$:

7. The Covariance

Theorem For the RV, X and Y...

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - E(X)E(Y)$$

Proof:

$$\begin{aligned}\text{Cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E(XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y) \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X\mu_Y.\end{aligned}$$

Because $E(X) = \mu_X$ and $E(Y) = \mu_Y$, it follows that

$$\text{Cov}(X,Y) = E(XY) - E(X)E(Y) = E(XY) - \mu_X\mu_Y.$$

7. The Covariance

If X and Y are independent, then...

$$Cov(X, Y) = 0$$

Yet the opposite is not true $Cov(X, Y) = 0$ does not imply independence

7. The Covariance

Example

Show that X and Y are dependent but have $Cov(X, Y) = 0$

		Y -1	Y 0	Y 1
	-1	1/16	3/16	1/16
X	0	3/16	0	3/16
	1	1/16	3/16	1/16

Calculation of marginal probabilities yields
$p_X(-1) = p_X(1) = 5/16 = p_Y(-1) = p_Y(1)$, and
$p_X(0) = 6/16 = p_Y(0)$. The value $p(0, 0) = 0$.

$$p(0, 0) \neq p_X(0)p_Y(0)$$

Hence, they are dependent.

7. The Covariance

Example

Show that X and Y are dependent but have $Cov(X, Y) = 0$

$$\begin{aligned}
E(XY) &= \sum_{\text{all } x} \sum_{\text{all } y} xyp(x, y) \\
&= (-1)(-1)\left(\frac{1}{16}\right) + (-1)(0)\left(\frac{3}{16}\right) + (-1)(1)\left(\frac{1}{16}\right) \\
&\quad + (0)(-1)\left(\frac{3}{16}\right) + (0)(0)(0) + (0)(1)\left(\frac{3}{16}\right) \\
&\quad + (1)(-1)\left(\frac{1}{16}\right) + (1)(0)\left(\frac{3}{16}\right) + (1)(1)\left(\frac{1}{16}\right) \\
&= \frac{1}{16} - \frac{1}{16} - \frac{1}{16} + \frac{1}{16} = 0.
\end{aligned}$$

7. The Covariance

Example

Show that X and Y are dependent but have $Cov(X, Y) = 0$
Looking at marginal probabilities we see that:

$$E(X) = E(Y) = 1 \times 5/16 + (-1) \times 5/16 = 0$$

Hence

$$\begin{aligned} \mathrm{Cov}(X, Y) &= E(XY) - E(X)E(Y) \\ &= 0 - 0(0) = 0. \end{aligned}$$

Contents

8. Linear Functions of Random Variables

Sometimes, we will encounter parameter estimators that are linear functions of random variables (measurements in a sample), $Y_1, Y_2, \ldots, Y_n$. If $a_1, a_2, \ldots, a_n$ are constants, we will need to find the expected value and variance of U_1:

$$U_1 = a_1Y_1 + a_2Y_2 + a_3Y_3 + \cdots + a_nY_n = \sum_{i=1}^{n} a_iY_i$$

We also may be interested in the covariance between two such linear combinations. Results that simplify the calculation of these quantities are summarized in the following theorem.

8. Linear Functions of Random Variables

Theorem Given:

- A set of random variables $Y_1, Y_2, \ldots, Y_n$ with expected values $E(Y_i) = \mu_i$.
- Constants $a_1, a_2, \ldots, a_n$

Define:

- U_1 as the linear combination of the Y_i variables: $U_1 = \sum_{i=1}^{n} a_i Y_i$.

8. Linear Functions of Random Variables

Then:

1. **Expected Value of U_1:**

$$E(U_1) = \sum_{i=1}^{n} a_i \mu_i$$

2. **Variance of U_1:**

$$V(U1) = \sum_{i=1}^{n} a_i^2 V(Y_i) + 2 \sum \sum_{1 \leq i < j \leq n} a_i a_j \mathrm{Cov}(Y_i, Y_j)$$

Here, the double sum covers all pairs of i, j where i is less than j.

8. Linear Functions of Random Variables

For example, if n=3...

1. **Expected Value of** U_1:

$$E(U_1) = a_1\mu_1 + a_2\mu_2 + a_3\mu_3$$

2. **Variance of** U_1:

$$V(U_1) = a_1^2V(Y_1) + a_2^2V(Y_2) + a_3^2V(Y_3)$$
$$+2(a_1a_2\text{Cov}(Y_1, Y_2) + a_1a_3\text{Cov}(Y_1, Y_3) + a_2a_3\text{Cov}(Y_2, Y_3))$$

8. Linear Functions of Random Variables

The sample mean

Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2$. (These variables may denote the outcomes of n independent trials of an experiment.) Define

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

and show that $E(\bar{Y}) = \mu$ and $V(\bar{Y}) = \frac{\sigma^2}{n}$.

8. Linear Functions of Random Variables

The sample mean

Solution: Notice that $\bar{Y}$ is a linear function of $Y_1, Y_2, \ldots, Y_n$ with all constants a_i equal to $\frac{1}{n}$. That is,

$$\bar{Y} = \frac{1}{n}Y_1 + \cdots + \frac{1}{n}Y_n$$

By the previous theorem (part 1),

$$E(\bar{Y}) = \sum_{i=1}^{n} a_i\mu_i = \sum_{i=1}^{n} a_i\mu = \mu\sum_{i=1}^{n} a_i = \mu\sum_{i=1}^{n}\frac{1}{n} = \frac{n\mu}{n} = \mu$$

8. Linear Functions of Random Variables

The sample mean

Solution:

By the theorem (part 2),

$$V(Y) = \sum_{i=1}^{n} a_i^2 V(Y_i) + 2 \sum_{i=1}^{n} \sum_{i<j} a_i a_j \mathsf{Cov}(Y_i, Y_j)$$

The covariance terms are all zero because the random variables are independent. Thus,

$$V(Y) = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma_i^2 = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$