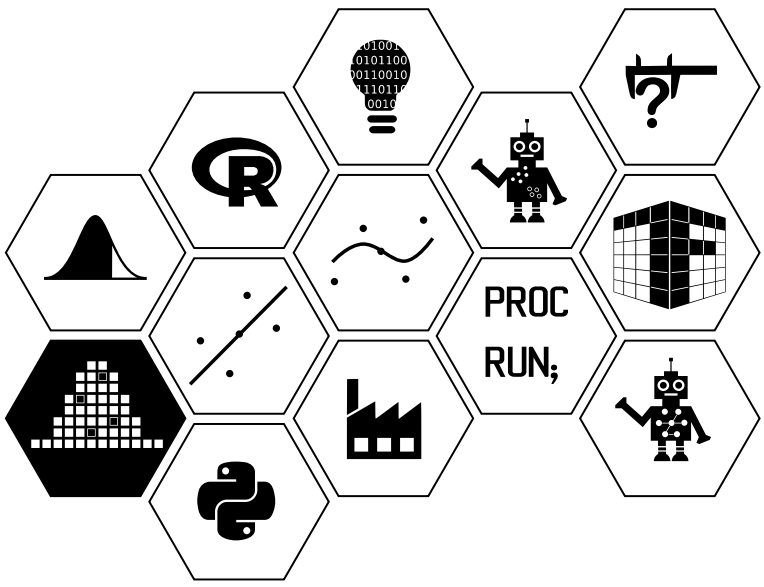
Probability and Sampling Fundamentals

Week 4: Bivariate Discrete Random Variables and the Multinomial Distribution



DATA ANALYTICS
GLASGOW

Bivariate discrete random variables and the multinomial distribution

This week's material is divided up into two separate but related parts. The first introduces joint properties of two **discrete** random variables before showing how these results can be extended to more than two random variables by introducing the concept of a **random vector**. The second part introduces by far the most commonly encountered standard distribution for a discrete random vector, the **multinomial distribution**.

Week 4 learning material aims

The material in week 4 covers:

- the joint bivariate probability mass function;
- obtaining marginal and conditional distributions for bivariate discrete random variables;
- independence of bivariate discrete random variables;
- extending these results to discrete random vectors;
- properties of the multinomial distribution.

Why multivariate probability matters

In week 3 we learned about single discrete random variables, called univariate random variables, and learned how to calculate some of their core properties such as the expectation and variance. However, often we are not just interested in how to describe a single random variable but instead in understanding how different variables relate to each other. For example we might be interested in whether

- alcoholics are more likely to develop dementia;
- sons grow to be taller than their mothers;
- courts in Florida are more likely to impose the death penalty on black defendants.

These questions involve the **joint** distribution of more than one random variable.

This week we will present some of the probability theory required to describe the joint properties of two **discrete** random variables, ideas which can also be extended to more than two random variables.

Bivariate discrete random variables - a motivating example

We will use the following example to help us define some of the important concepts for this week's material.

Example 1

Motivating example

Tennis match

Serena and Venus are two professional tennis players. When they play a match against each other, the winner of the match is the first to win two sets. To record the result of one match, the following discrete random variables can be used:

$$X = \{\text{the number of sets won by Serena}\},$$
$$Y = \{\text{the number of sets won by Venus}\}.$$

Joint range space

The possible outcomes of any match, (X, Y), in this example are $(0, 2), (1, 2), (2, 0), (2, 1)$, where, for example, $(2, 0)$ corresponds to Serena winning the match two sets to zero and $(0, 2)$ corresponds to

Venus winning the match by two sets to zero. We can use this to find the **joint range space** of (X, Y), $R_{XY} = \{(0, 2), (1, 2), (2, 0), (2, 1)\}$. This leads us to the formal definition.

Definition 1

Joint range space

Let X and Y be two random variables. The **joint range space** R_{XY} of X and Y is the set of all pairs of values (X, Y) can take. Or mathematically,

$$R_{XY} = \{(X(s), Y(s)) : s \in S\}.$$

where s are elements of the sample space S.

Note: Remember from last week that a **discrete** joint distribution has a range space that is finite or countable.

Joint probability mass function

Example 2

Motivating example continued

In Example 1, we know that Serena is ranked much higher in the player rankings than Venus and it is believed when they play that Serena has a 70% chance of winning each set in a match, independent of the previous set. We can use this information to find the probability of an event such as Serena winning the match 2-1. To do that we need to compute $p_{XY}(2, 1) = P(X = 2, Y = 1)$.

We can calculate all of the probabilities of the form $p_{XY}(x, y) = P(X = x, Y = y)$ by considering the distribution of X, the number of sets won by Serena, only.

We will assume that Serena wins a set with probability $\theta = 0.7$, independently of the other sets. If we knew the number of sets n played, we could write

$$X \sim \text{Bin}(n, \theta).$$

We can exploit this similarity to the Binomial distribution when we calculate the probabilities $p_{XY}(x, y) = P(X = x, Y = y)$ of the different outcomes.

Firstly, $p_{XY}(x,y) = 0$ when (x,y) lie outside the joint range space, $R_{XY} = \{(0,2),(1,2),(2,0),(2,1)\}$. For example, the probability of Serena winning 3 sets and Venus winning 1, $p_{XY}(3,1)$, is equal to 0 since the game is over as soon as one player wins 2 sets.

The remaining joint probabilities which are not equal to 0 are calculated as follows.

$$
\begin{aligned}
p_{XY}(2,0) &= P(\text{S wins first set and second set}) \\
&= \binom{2}{2}\theta^2(1-\theta)^0 \\
&= \theta^2 \\
&= 0.7^2 = 0.490
\end{aligned}
$$

$$
\begin{aligned}
p_{XY}(0,2) &= P(\text{S loses first set and second set}) \\
&= \binom{2}{0}\theta^0(1-\theta)^2 \\
&= (1-\theta)^2 \\
&= 0.3^2 = 0.09
\end{aligned}
$$

$$
\begin{aligned}
p_{XY}(2,1) &= P(\text{S wins one of the first two sets and S wins the third set}) \\
&= P(\text{S wins one of the first two sets}) \cdot P(\text{S wins the third set}) \qquad (\text{independence}) \\
&= \binom{2}{1}\theta^1(1-\theta)^1 \cdot \binom{1}{1}\theta^1(1-\theta)^0 \\
&= 2\theta(1-\theta)\cdot\theta \\
&= 2\theta^2(1-\theta) \\
&= 2 \cdot 0.7^2 \cdot 0.3 = 0.294
\end{aligned}
$$

$$
\begin{aligned}
p_{XY}(1,2) &= P(\text{S wins one of the first two sets and S loses the third set}) \\
&= P(\text{S wins one of the first two sets}) \cdot P(\text{S loses the third set}) \qquad (\text{independence}) \\
&= \binom{2}{1}\theta^1(1-\theta)^1 \cdot \binom{1}{0}\theta^0(1-\theta)^1 \\
&= 2\theta(1-\theta)\cdot(1-\theta) \\
&= 2\theta(1-\theta)^2 \\
&= 2 \cdot 0.7 \cdot 0.3^2 = 0.126
\end{aligned}
$$

These probabilities are shown in the table below. You may notice that all of the probabilities are non-negative and smaller than 1. Furthermore, the sum of all of the joint probabilities is equal to 1.

$p_{XY}(x,y)$			x	
		0	1	2
	0	0	0	0.490
y	1	0	0	0.294
	2	0.090	0.126	0

The probability mass function takes two inputs (x and y) and returns one value ($p_{XY}(x,y)$), so we can visualise in a 3D plot.

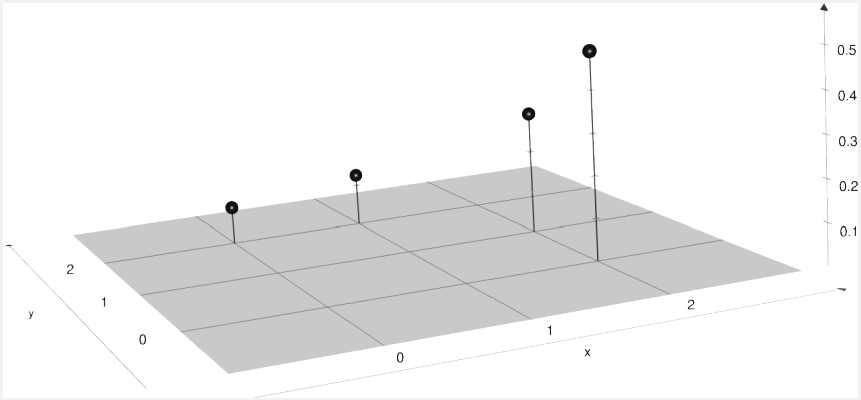


Figure 1

However, a more informative plot is obtained when showing the probability as the size of a dot is a plot of the joint range space.

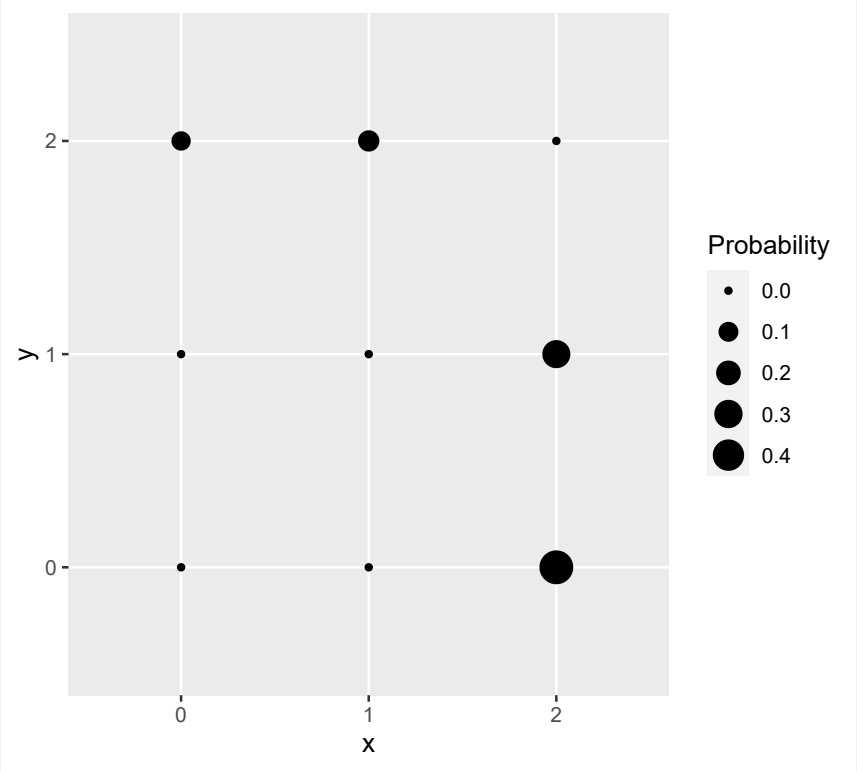


Figure 2

The probability, $p_{XY}(x, y)$, is called the **joint bivariate probability mass function** and is formally defined below.

Definition 2

Joint probability mass function

Let (X, Y) be a pair of discrete random variables. The function

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

is called the **joint probability mass function (p.m.f.)** of X and Y.

Just like the univariate probability mass function, the joint probability mass function must satisfy the following two conditions:

1. $0 \leq p_{XY}(x, y) \leq 1$, for every $(x, y) \in \mathbb{R}^2$; each joint probability must be between 0 and 1;
2. $\sum_{(x,y) \in R_{XY}} p_{XY}(x, y) = 1$; the sum of the probabilities is equal to 1.

The joint probability mass function might be given by a table like in Example 1 above, or might be given in the form of an equation like

$$p_{XY}(x,y) = \begin{cases} \frac{n!}{y!(x-y)!(n-x)!}\theta^x(1-\theta)^{n-x}\phi^y(1-\phi)^{x-y} & \text{for } 0 \leq y \leq x \leq n \\ 0 & \text{otherwise.} \end{cases}$$

We will mostly focus on the case when the joint probability mass function is given by a table.

We can calculate the probabilities of composite events by summing the probability mass function over the set corresponding to that event, just in the same way as we have done last week for univariate ("one-dimensional") random variables.

Example 3

Motivating example continued

Use the joint probability mass function to find the probability that both players win at least one set.

Answer:

$$\begin{aligned} P(\text{both players win at least one set}) &= p_{XY}(X=1,Y=2) + p_{XY}(X=2,Y=1) \\ &= p_{XY}(1,2) + p_{XY}(2,1) \\ &= 0.126 + 0.294 \\ &= 0.42 \end{aligned}$$

Marginal distributions

Say we are interested in calculating the probability that Serena wins 2 sets, regardless of how many sets Venus wins. To calculate this we can sum up each of the joint probabilities where $X = 2$ (i.e. $p_X(x=2)$). If we look at the joint probability mass function above, this is equivalent to summing the last column in the table. You can think of this as $p_X(X=2) = p_X(X=2,Y=0) + p_X(X=2,Y=1)$. This makes sense since the probability of Serena winning the game can happen in two different ways.

In general, $p_X(x)$ is the sum of the appropriate **column** of the joint probability mass function (assuming that the values of X define the columns). The probability, $p_X(x)$ is known as the **marginal probability mass function** of X and is the same probability distribution that would have been defined if X had been the only piece of information recorded from the experiment. For example, if we had only collected information about the number of sets won by Serena in this tennis match. The marginal probability

mass function of Y, $p_Y(y)$, is calculated similarly however now the appropriate **row** of the joined probability mass function is summed.

Example 4

Motivating example continued

The following table shows these marginal distributions for Example 1.

$p_{XY}(x,y)$			x		
		0	1	2	$p_Y(y)$
	0	0	0	0.490	0.490
y	1	0	0	0.294	0.294
	2	0.090	0.126	0	0.216
	$p_X(x)$	0.090	0.126	0.784	1

So to answer our question, the probability of Serena winning 2 sets (irrespective of how many Venus wins) is $p_X(2) = 0.784$.

These marginal distributions can be calculated more formally as follows:

$$\begin{aligned} p_X(0) &= p_{XY}(0,0) + p_{XY}(0,1) + p_{XY}(0,2) \\ &= 0 + 0 + 0.090 \\ &= 0.090 \end{aligned}$$

$$\begin{aligned} p_Y(0) &= p_{XY}(0,0) + p_{XY}(1,0) + p_{XY}(2,0) \\ &= 0 + 0 + 0.490 \\ &= 0.490 \end{aligned}$$

$$\begin{aligned} p_X(1) &= p_{XY}(1,0) + p_{XY}(1,1) + p_{XY}(1,2) \\ &= 0 + 0 + 0.126 \\ &= 0.126 \end{aligned}$$

$$\begin{aligned} p_Y(1) &= p_{XY}(0,1) + p_{XY}(1,1) + p_{XY}(2,1) \\ &= 0 + 0 + 0.294 \\ &= 0.294 \end{aligned}$$

$$\begin{aligned} p_X(2) &= p_{XY}(2,0) + p_{XY}(2,1) + p_{XY}(2,2) \\ &= 0.490 + 0.294 + 0 \\ &= 0.784 \end{aligned}$$

$$\begin{aligned} p_Y(2) &= p_{XY}(0,2) + p_{XY}(1,2) + p_{XY}(2,2) \\ &= 0.090 + 0.126 + 0 \\ &= 0.216 \end{aligned}$$

This leads us to the formal definition of a marginal probability mass function.

Definition 3

Marginal probability mass function

We can obtain the **marginal probability mass function** $p_X(x)$ and $p_Y(y)$ from their joint probability mass function by

$$p_X(x) = \sum_{y:(x,y)\in R_{XY}} p_{XY}(x,y)$$

$$p_Y(y) = \sum_{x:(x,y)\in R_{XY}} p_{XY}(x,y)$$

Note: The summation, $\sum_{y:(x,y)\in R_{XY}}$, simply means sum over all possible values of y and similarly, $\sum_{x:(x,y)\in R_{XY}}$, means sum over all possible values of x. In other words to find the marginal probability mass function of one random variable we sum over the possible values of the **other** random variable.

This video provides a detailed explanation on constructing bivariate probability mass functions and how to use these to calculate marginal distributions.

Video

Bivariate discrete random variables

Duration 3:51

Expectation and variance

We can find the expectation and variance of X and Y from the marginal probability mass function as usual:

$$E(X) = \sum_{x\in R_X} x p_X(x) = 0\cdot 0.090 + 1\cdot 0.126 + 2\cdot 0.784 = 1.694$$

$$E(X^2) = \sum_{x\in R_X} x^2 p_X(x) = 0^2\cdot 0.090 + 1^2\cdot 0.126 + 2^2\cdot 0.784 = 3.262$$

$$\text{Var}(X) = E[X^2] - [E(X)]^2 = 3.262 - [1.694]^2 = 0.392$$

and

$$E(Y) = \sum_{y \in R_Y} y p_Y(y) = 0 \cdot 0.490 + 1 \cdot 0.294 + 2 \cdot 0.216 = 0.726$$

$$E(Y^2) = \sum_{y \in R_Y} y p_Y(y) = 0^2 \cdot 0.490 + 1^2 \cdot 0.294 + 2^2 \cdot 0.216 = 1.158$$

$$\text{Var}(Y) = E[Y^2] - [E(Y)]^2 = 1.158 - [0.726]^2 = 0.631$$

In Example 1, the expected number of sets that Serena will win is 1.694 with a variance of 0.392. Whereas the expected number of sets that Venus will win is 0.726 with a variance of 0.631.

Note: Marginal expectations and variances summarise properties (location and spread) of the individual random variables, they do not yet summarise relationships between them.

Using the joint probability mass function it is possible to find the **expected value** of any real function, $g(X, Y)$, of X and Y. This allows us to summarise relationships between random variables by calculating the **covariance** and **correlation**.

Definition 4

Expected value

Let (X, Y) be a pair of discrete random variables and $g(X, Y)$ be any real-valued function of X and Y. Then, if it exists, the **expected value** of $g(X, Y)$ is defined to be

$$E[g(X, Y)] = \sum_{(x,y) \in R_{XY}} \sum g(x, y) \cdot p_{XY}(x, y)$$

The marginal expectation $E(X)$ can be obtained by setting $g(X, Y) = X$ and the marginal expectation $E(Y)$ can be obtained by setting $g(X, Y) = Y$.

Example 5

Motivating example continued

In Example 1, if we let $g(X,Y) = XY$ then $E[g(X,Y)] = E(XY)$ is calculated as follows:

$$\begin{aligned} E(XY) &= \sum_{(x,y)\in R_{XY}} \sum xy \cdot p_{XY}(x,y) \\ &= (0 \times 2 \times 0.090) + (1 \times 2 \times 0.126) \\ &+ (2 \times 0 \times 0.490) + (2 \times 1 \times 0.294) \\ &= 0 + 0.252 + 0 + 0.588 \\ &= 0.84 \end{aligned}$$

Covariance

It is often of interest to look at relationships between pairs of variables (e.g. the relationship between alcoholics and dementia). **Covariance** allows us to summarise linear relationships between two random variables.

Definition 5

Covariance

The **covariance** between two random variables X and Y is defined as

$$\text{Cov}(X,Y) = E((X - E(X))(Y - E(Y)))$$

- A **positive** value of the covariance indicates that there is a **positive linear relationship** between two random variables; i.e., higher values of one tend to be observed along with higher values of the other.
- A **negative** value of the covariance indicates that there is a **negative linear relationship** between two random variables; i.e., higher values of one tend to be observed along with lower values of the other.

Re-writing the definition of the covariance yields the alternative formula

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y).$$

Other important properties of the covariance are:

- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$ (order of arguments does not matter),
- $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$ (covariance between a random variable and itself is its variance),
- $-\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)} \leq \mathrm{Cov}(X,Y) \leq \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$ (range of possible values of covariance depends on variances).

Task 1 (Optional)

Show that

a) $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$,

b) $\mathrm{Cov}(Y,X) = \mathrm{Cov}(X,Y)$,

c) $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

Example 6

Motiating example continued

In Example 1 the covariance between X, the number of sets won by Serena, and Y, the number of sets won by Venus, is calculated as follows.

We have previously calculated,

$$E(X) = 1.694, \quad E(Y) = 0.726, \quad E(XY) = 0.84.$$

Then,

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E(X,Y) - E(X)E(Y)\\ &= 0.840 - 1.694 \cdot 0.726\\ &= 0.840 - 1.230\\ &= -0.390\end{aligned}$$

Here, a negative covariance makes sense since if one player scores highly, the other player must have scored a lower number of sets.

Correlation

Although a positive covariance indicates a positive linear association and a negative covariance indicates a negative linear association, it is difficult to interpret the actual number. For example, a covariance of 2 may indicate a **weak** or **strong** positive relationship depending on the variability in the data. Instead, the **correlation**, which is a measure of association between two variables based on the covariance, can be calculated which has a much easier interpretation since it has an absolute meaning.

Definition 6

Correlation

Let X and Y be two random variables with non-zero variances. Then the **correlation** of X and Y is

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

As we can see from the formula of covariance, it assumes the units from the product of the units of the two variables. On the other hand, correlation is dimensionless. It is a unit-free measure of the relationship between variables. This is because we divide the value of covariance by the product of standard deviations which have the same units. If all the values of the given variable are multiplied by a constant and all the values of another variable are multiplied, by a similar or different constant, then the value of covariance also changes. However, on doing the same, the value of correlation is not influenced by the change in scale of the values. Another difference between covariance and correlation is the range of values that they can take. The key property of correlation is that correlation coefficients lie between -1 and 1, i.e.

$$-1 \leq \rho(X,Y) \leq 1,$$

whereas covariance can take any value between $-\infty$ and $+\infty$, as it depends on $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$.

The sign of the correlation has the same interpretation as the sign of the covariance. Positive correlation indicates a positive linear relationship and negative correlation indicates a negative linear relationship.

However, now the value of the correlation informs us how strong the linear relationship between X and Y is.

- If $\rho = 0$ there is no linear relationship. In this case we say that X and Y are *uncorrelated*.
- If $\rho = \pm 1$ there is a perfect linear relationship.

The figure below shows six joint probability mass functions with identical marginal distributions but different correlations.

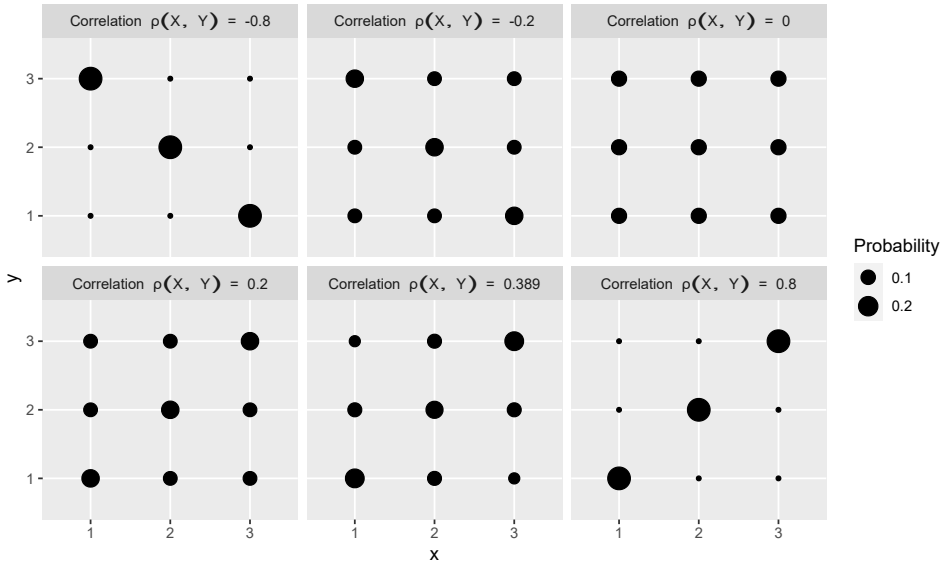


Figure 3

Example 7

Motivating example continued

In Example 1 the correlation between X and Y is calculated as follows.

We have previously calculated,

$$\text{Var}(X) = 0.392, \quad \text{Var}(Y) = 0.631, \quad \text{Cov}(X, Y) = -0.390.$$

$$\begin{aligned}\rho(X,Y) &= \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \\ &= \frac{-0.390}{\sqrt{0.392 \times 0.631}} \\ &= \frac{-0.390}{0.497} \\ &= -0.785\end{aligned}$$

Hence, there is a reasonably **strong, negative** linear relationship between the number of sets won by Serena (X) and the number of sets won by Venus (Y).

Task 2

In an experiment, adult female subjects are to have their distance vision tested in both eyes. A subject will be given a grade of 1, 2 or 3 for each eye (where 1 indicates the best sight). For every patient the experiment returns a pair of numbers (X, Y) where

$$X = \{\text{grade of distance vision in right eye}\}$$

$$Y = \{\text{grade of distance vision in left eye}\}$$

The joint probability mass function of (X, Y) is given by the table below.

$p_{XY}(x,y)$			x		
		1	2	3	$p_Y(y)$
	1	0.2	0.03	0.02	
y	2	0.04	0.3	0.06	
	3	0.01	0.07	0.27	
	$p_X(x)$				1

1. Complete the table by computing the marginal distribution of X and Y.
2. Calculate the probability that a subject gets

(a) the best grade in her right eye and the worst grade in her left eye;
(b) the best grade for one eye and the worst grade for the other eye;

(c) the same grade for both eyes;
(d) a better grade for her right eye than for her left eye.

3. Calculate the expectation and variance of the grades in the left and right eyes.

4. Calculate the correlation between the left and right eyes.

Conditional distributions

As discussed in week 2, conditional probability allows us to update our beliefs about uncertain events when we get new information. In Example 1, if we know that if $X = 0$ then Y must equal 2 (i.e. Venus must have won the match 2 sets to 0):

$$P(Y = 2|X = 0) = 1.$$

More generally, if we know that X takes the value x, then the distribution of Y is given by the **conditional distribution** of Y given that $X = x$.

Definition 7

Conditional probability mass function

Let (X, Y) be a pair of discrete random variables. Then by the definition of conditional probability (Definition 1, Week 2),

$$P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{p_{XY}(x, y)}{p_X(x)}$$
$$\text{for } x \in R_X \text{ and } y \text{ such that } (x, y) \in R_{XY}$$

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p_{XY}(x, y)}{p_Y(y)}$$
$$\text{for } y \in R_Y \text{ and } x \text{ such that } (x, y) \in R_{XY}$$

Example 8

Motivating example continued

In Example 1, the conditional p.m.f. of Y given $X = 2$, $p_{Y|X}(y|2)$, is

$$
\begin{aligned}
P_{Y|X}(0|2) &= \frac{p_{XY}(2,0)}{p_X(2)} = \frac{0.490}{0.784} = 0.625, \\
P_{Y|X}(1|2) &= \frac{p_{XY}(2,1)}{p_X(2)} = \frac{0.294}{0.784} = 0.375, \\
P_{Y|X}(2|2) &= \frac{p_{XY}(2,2)}{p_X(2)} = \frac{0}{0.784} = 0.
\end{aligned}
$$

Note:

- The terminology $P_{Y|X}(0|2)$ denotes the conditional probability of Y **given** X when $Y = 0$ and $X = 2$. In other words, what is the probability that Y will be 0 given we know that X is equal to 2?
- We technically only need to calculate two probabilities here, $P_{Y|X}(0|2)$ and $P_{Y|X}(1|2)$ since the question asks for $p_{Y|X}(y|2)$, the conditional probability of Y given $X = 2$. In Example 1, when $X = 2$ we know that Y can only take two values, 0 or 1 since it is impossible for Serena and Venus to both win two sets in a match.

This video introduces the concepts of covariance, correlation, and marginal distributions using Example 1.

Video

Bivariate discrete random variables continued

Duration 12:19

Task 3

Consider again the joint p.m.f. from Task 2.

$$X = \{\text{grade of distance vision in right eye}\}$$

$$Y = \{\text{grade of distance vision in left eye}\}$$

The joint probability mass function of (X, Y) is given by the table below.

$p_{XY}(x,y)$			x		
		1	2	3	$p_Y(y)$
	1	0.2	0.03	0.02	0.25
y	2	0.04	0.3	0.06	0.40
	3	0.01	0.07	0.27	0.35
	$p_X(x)$	0.25	0.40	0.35	1

Find the conditional p.m.f. $p_{X|Y}(x|1)$.

Independence

In week 2 we learned about the important relationship between conditional probability and independence. Similarly we can relate the concepts of independence and conditional distributions. Let's look again at Example 1.

Example 9

Motivating example continued

Suppose we wanted to know if the number of sets won by Serena (X) was independent of the number of sets won by Venus (Y). We have in fact already partially answered this question when we calculated the conditional p.m.f. of Y given $X = 2$ ($p_{Y|X}(y|2)$). We found

$$P_{Y|X}(0|2) = \frac{p_{XY}(2,0)}{p_X(2)} = \frac{0.490}{0.784} = 0.625,$$
$$P_{Y|X}(1|2) = \frac{p_{XY}(2,1)}{p_X(2)} = \frac{0.294}{0.784} = 0.375,$$
$$P_{Y|X}(2|2) = \frac{p_{XY}(2,2)}{p_X(2)} = \frac{0}{0.784} = 0.$$

Intuitively, if X and Y were independent here then the probability of one occurring would not impact the probability of the other occurring. In other words, the **conditional** p.m.f. would be equal to the **marginal** p.m.f. For Example 1, let's compare the conditional p.m.f. of $Y = 0$ given $X = 2$,

$$p_{Y|X}(0|2) = 0.625,$$

to the marginal p.m.f. of $Y = 0$,

$$p_Y(0) = 0.490.$$

Since these two probabilities are not equal to one another we know that X and Y are **not independent**.

We will now state what we have used in the example in general. In week 2 we also learned that two events A and B are independent if

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

The same approach can be used to define independence of **random variables**.

Definition 8

Independence of random variables

Two random variables X and Y are independent if any event relating to one of the random variables is independent of any event relating to the other random variable.

One can show that this corresponds to the joint p.m.f. factorising into the two marginal p.m.f.'s (i.e. $P(A \cap B) = P(A)P(B)$).

Proposition 1

Let X and Y be two discrete random variables, then the following four statements are equivalent.

(i) X and Y are independent.

(ii) The joint probability mass function of X and Y is the product of the marginal probability mass functions, i.e.

$$p_{XY}(x,y) = p_X(x)p_Y(y) \text{ for all } x \text{ and } y.$$

(iii) The conditional distribution of X given $Y = y$ does not depend on y, i.e.

$$p_{X|Y}(x|y) = p_X(x) \quad \text{for all } x \text{ and } y.$$

(iv) The conditional distribution of Y given $X = x$ does not depend on x, i.e.

$$p_{Y|X}(y|x) = p_Y(y) \quad \text{for all } x \text{ and } y.$$

Part (ii) of the proposition tells us that two discrete random variables X and Y are independent if and only if

- the joint probability mass function is the product of the marginal probability mass functions of the joint range space, i.e.

$$p_{XY}(x,y) = p_X(x)p_Y(y) \text{ for all } x \in R_X \text{ and } y \in R_Y, \quad \text{and}$$

- the joint range space R_{XY} is the **Cartesian product** of the range spaces of X and Y, i.e. $R_{XY} = R_X \times R_Y$. What this means is that every possible combination (x,y) of the outcomes $x \in R_X$ of X and the outcomes $y \in R_Y$ of Y must be able to occur.

The latter can provide a quick way of showing that two random variables are not independent.

Note: To see where the second bullet point comes from, consider any $x \in R_X$ (any value of x in the range space of X) and any $y \in R_Y$ (any value of y in the range space of Y) such that $(x,y) \notin R_{XY}$ (the values (x,y) are not in the joint range space of X and Y). Then, by definition $p_X(x)$ and $p_Y(y)$ are greater than 0 but $p_{XY}(x,y) = 0$, which means $p_{XY}(x,y) \neq p_X(x)p_Y(y)$.

Example 10

Motivating example continued

In Example 1, it is possible that Serena wins two sets (i.e. $2 \in R_X$) and it is equally possible that Venus wins to sets (i.e. $2 \in R_Y$), but they cannot both win two sets (i.e. $(2,2) \notin R_{XY}$), hence X and Y cannot be independent.

Another way to show that X and Y are not independent is to find one combination of (x,y) where

$$p_{XY}(x,y) \neq p_X(x) \cdot p_Y(y).$$

Example 11

Motivating example continued

For example, in Example 1, when $x = 0$ and $y = 0$

$$\begin{aligned} p_{XY}(x,y) &= 0, \\ p_X(x)p_Y(y) &= 0.090 \cdot 0.490 = 0.004, \\ p_{XY}(x,y) &\neq p_X(x)p_Y(y). \end{aligned}$$

thus we have again shown formally that X and Y are not independent.

We could have also used the argument we have used above that

$$p_{Y|X}(0|2) = 0.625 \neq 0.490 = p_Y(y).$$

Note: to show that X and Y are not independent, it is sufficient to find one combination of (x,y) where

$$p_{XY}(x,y) \neq p_X(x)p_Y(y) \quad \text{or} \quad p_{X|Y}(x|y) \neq p_X(x) \quad \text{or} \quad p_{Y|X}(y|x) \neq p_Y(y).$$

Note that we can always compute the joint probability mass function as a product of the marginal probability mass function of one of the random variables and the conditional probability mass function of the other, i.e.

$$
\begin{aligned}
p_{XY}(x,y) &= p_X(x)p_{Y|X}(y|x) \\
&= p_Y(y)p_{X|Y}(x|y),
\end{aligned}
$$

However, if X and Y are independent, then $p_{X|Y}(x|y) = p_X(x)$ and $p_{Y|X}(y|x) = p_Y(y)$ and both lines of the above equation just correspond to $p_{XY}(x,y) = p_X(x)p_Y(y)$.

Example 12

Independence

Rolling a pair of dice

Suppose we roll a pair of dice. Denote by X and Y the score of the first and second dice, respectively. Are the random variables X and Y independent?

Answer:

Firstly, the joint range space is

$$
\begin{aligned}
R_{XY} &= \{(x,y) : x = 1,\ldots,6, y = 1,\ldots,6\} \\
&= \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,6)\} \\
&= \{1,\ldots,6\} \times \{1,\ldots,6\} \\
&= R_X \times R_Y
\end{aligned}
$$

and for $x \in R_X$ and $y \in R_Y$,

$$
\begin{aligned}
p_{XY}(x,y) &= \frac{1}{36}, \\
p_X(x)p_Y(y) &= \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}, \\
p_{XY}(x,y) &= p_X(x)p_Y(y),
\end{aligned}
$$

thus X and Y are independent.

Task 4

Consider again the joint p.m.f. from Task 1.

$$X = \{\text{grade of distance vision in right eye}\}$$

$$Y = \{\text{grade of distance vision in left eye}\}$$

The joint probability mass function of (X, Y) is given by the table below.

$p_{XY}(x,y)$			x		
		1	2	3	$p_Y(y)$
	1	0.2	0.03	0.02	0.25
y	2	0.04	0.3	0.06	0.40
	3	0.01	0.07	0.27	0.35
	$p_X(x)$	0.25	0.40	0.35	1

Are the random variables X and Y independent?

Independence and Correlation

We finally look at the relationship between correlation and independence. Both tell us something about the dependency between X and Y. However correlation only measures *linear* dependency, whereas independence can capture any type of dependency.

Proposition 2

Let X and Y be two independent random variables, then $\rho(X, Y) = 0$, i.e. X and Y are uncorrelated.

What this proposition means is that two random variables which are independent are also uncorrelated. Note that the converse is not true, i.e. two random variables can be uncorrelated, but still not be independent. This is the case if the relationship between the two random variables is entirely nonlinear, as the example below shows.

Example 13

Consider a discrete random variable X with probability mass function

$$p_X(x) = \begin{cases} \frac{1}{2} & \text{if} x = 0 \\ \frac{1}{4} & \text{if } x = -1 \text{ or } 1 \end{cases}$$

and define $Y = X^2$, i.e. X and Y are deterministically linked. The joint probability mass function is then given by the table below.

$p_{XY}(x,y)$			x		
		-1	0	1	$p_Y(y)$
	0	0	0.5	0	0.5
y	1	0.25	0	0.25	0.5
	$p_X(x)$	0.25	0.5	0.25	1

X and Y are not independent, as for example

$$p_{XY}(0,1) = 0 \neq \frac{1}{2} \times \frac{1}{2} = p_X(0)p_Y(1).$$

However,

$$
\begin{aligned}
E(X) &= \sum_{x \in R_X} x p_X(x) = -1 \times p_X(-1) + 0 \times p_X(0) + 1 \times p_X(1) \\
&= -1 \times \frac{1}{4} + 0 \times \frac{1}{2} + 1 \times \frac{1}{4} = 0 \\
E(Y) &= \sum_{y \in R_Y} y p_Y(y) = 0 \times p_Y(0) + 1 \times p_Y(1) \\
&= 0 \times \frac{1}{2} + 1 \times \frac{1}{2} = \frac{1}{2} \\
E(XY) &= \sum_{(x,y) \in R_{XY}} xy p_{XY}(x,y) \\
&= -1 \times 0 \times p_{XY}(-1,0) + 0 \times 0 \times p_{XY}(0,0) + \ldots + 1 \times 1 \times p_{XY}(1,1) \\
&= -1 \times 1\frac{1}{4} + 0 \times 0\frac{1}{2} + 1 \times 1\frac{1}{4} = 0
\end{aligned}
$$

Hence,

$$
\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = 0 - 0 \times \frac{1}{2} = 0,
$$

i.e. X and Y are uncorrelated, despite not being independent.

Supplement 1

The Datasaurus Dozen

As stated above, two variables being uncorrelated does not mean that there is no relationship between the variables (or that the variables are independent). This website visualises this concept quite nicely by plotting 13 fictitious datasets all with the same correlation but with vastly different appearances.

Task 5 (Optional)

Show that if X and Y are independent, then $E(XY) = E(X)E(Y)$ and thus $\rho(X,Y) = 0$.

Supplement 2

Non-tabulated diustributions

The joint, marginal or conditional distributions of X and Y are not always given by a table as in the examples we have seen so far. In this example this will not be the case.

Imagine that there are on average λ passengers on a train. If we denote by X the number of passengers on the train, it seems reasonable to assume that X has a Poisson distribution with mean λ, i.e. for $x = 0, 1, 2, \ldots$

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Now assume that every passenger has a valid ticket with probability θ, independently of other passengers. So if there are x passengers on the train, then the number of passengers with a ticket has a binomial distribution with parameters x and θ. Note that we are looking at the number of passengers with a ticket given that we know how many passengers are on the train, in order words, we are looking at a conditional distribution.

In more mathematical terms,

$$p_{Y|X}(y|x) = \binom{x}{y}\theta^y(1-\theta)^{x-y}$$

for $y = 0, 1, \ldots, x$.

We can now find the joint distribution of X and Y.

$$\begin{aligned} p_{XY}(x,y) &= p_X(x)p_{Y|X}(y|x) \\ &= \frac{e^{-\lambda}\lambda^x}{x!}\underbrace{\binom{x}{y}}_{\frac{x!}{y!(x-y)!}}\theta^y(1-\theta)^{x-y} \\ &= \frac{e^{-\lambda}\lambda^x\theta^y(1-\theta)^{x-y}}{y!(x-y)!} \end{aligned}$$

for $0 \leq y \leq x$.

What is the distribution of the number of passengers without a ticket if we do *not* know the total number of passengers on the train? In other words, what is the marginal distribution of Y?

$$
\begin{aligned}
p_Y(y) &= \sum_{x:\ (x,y)\in R_{XY}} p_{XY}(x,y) \\
&= \sum_{x=y}^{+\infty} \frac{e^{-\lambda}\lambda^x\theta^y(1-\theta)^{x-y}}{y!(x-y)!}
\end{aligned}
$$

The sum starts at y, which makes calculations a little difficult, we will now rewrite the sum using $z = x - y$, i.e. $x - y$ becomes z and x becomes $y + z$. This gives

$$
\begin{aligned}
p_Y(y) &= \sum_{z=0}^{+\infty} \frac{e^{-\lambda}\lambda^{y+z}\theta^y(1-\theta)^z}{y!z!} \\
&= \frac{e^{-\lambda}}{y!}(\lambda\theta)^y \underbrace{\sum_{z=0}^{+\infty} \frac{(\lambda(1-\theta))^z}{z!}}_{=e^{\lambda(1-\theta)}} \\
&= \frac{\overbrace{e^{-\lambda+\lambda(1-\theta)}}^{=e^{-\theta\lambda}}(\lambda\theta)^y}{y!} \\
&= \frac{e^{-\theta\lambda}(\theta\lambda)^y}{y!}
\end{aligned}
$$

In the third-last line we have used the Taylor series expansion of the exponential function, i.e. $e^a = \sum_{z=0}^{\infty} \frac{a^z}{z!}$.

The probability mass function of Y is however nothing other than the probability mass function of a Poisson distribution with rate $\theta\lambda$, so

$$
Y \sim \text{Pois}(\theta\lambda).
$$

If we also introduce the number of passengers *without* a ticket, Z, then $Z = X - Y$, i.e. the number of passengers without a ticket (Z) is the number of passengers on the train (X) minus the number of passengers with a ticket (Y).

One can then show that $Z \sim \text{Pois}((1-\theta)\lambda)$ and show the slightly surprising fact that Y and Z are independent (if we do not know how many passengers are on the train), despite Z being defined as $Z = X - Y$.

Random vectors

So far we have only considered **bivariate discrete** random variables however all of the results that have been stated so far can be generalised to **random vectors** of any dimension. The following example will extend some of the results we have seen above to a random vector containing three random variables.

Example 14

Random vector

Rolling three dice

Suppose we roll three dice. Let

$$X = \{\text{score of first dice}\},$$

$$Y = \{\text{score of second dice}\},$$

$$Z = \{\text{score of third dice}\}.$$

Joint range space

The joint range space R_{XYZ} of X, Y, and Z is the set of all triplets of values (X, Y, Z) can take. For this example, there are $6^3 = 216$ possible outcomes from rolling the three dice which are $(1,1,1), (1,1,2), \ldots, (1,1,6), (1,2,1), \ldots, (6,6,6)$.

Probability mass function for discrete random vector

We now have 3 discrete random variables so the joint probability mass function is simply

$$p_{XYZ}(x,y,z) = P(X = x, Y = y, Z = z).$$

For example the joint probability for the outcome that the score for each dice is 6, is

$$p_{XYZ}(x,y,z) = P(X = 6, Y = 6, Z = 6) = \frac{1}{216}.$$

Independence

The joint probability above was calculated using the knowledge that the random variables X, Y, and Z are independent, since the score of one dice will not have any impact on the score on any of the other dice. Therefore we can use the multiplication rule for independent events to calculate

$$\begin{aligned} p_{XYZ}(X = 6, Y = 6, Z = 6) &= p_X(X = 6) p_Y(Y = 6) p_Z(Z = 6), \\ &= \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6}, \\ &= \frac{1}{216}. \end{aligned}$$

Or, more generally

$$p_{XYZ}(x, y, z) = p_X(x)p_Y(y)p_Z(z).$$

Marginal probability mass functions

Sometimes we are not interested in the entire joint distribution of a random vector, but only in the distribution of a subset of the random variables in a random vector. Just like we have seen above for bivariate discrete random variables (**Definition 3**) we compute marginal distributions by summing over the variables we are not interested in. In this example, say we wanted to calculate the marginal distribution of the score of the first dice, $p_X(x)$. To do this we simply sum over the possible values of the **other** random variables, Y and Z.

$$p_X(x) = \sum_{y,z:(x,y,z)\in R_{XYZ}} p_{XYZ}(x, y, z).$$

We will now introduce one of the most commonly used multivariate discrete distributions, the **multinomial distribution**.

start here

Multinomial distribution

The **multinomial** distribution is a generalisation of the binomial distribution which was introduced in week 3. As a reminder, we obtain a binomial distribution when carrying out n independent Bernoulli trials, which are trials with only **two** possible outcomes ("success" with probability θ and "failure" with probability $1 - \theta$).

Now suppose that each trial can yield k different outcomes, rather than just two different outcomes. Examples include...

- Suppose we record for every student in a university course whether they pass at the degree exam, pass at the resit, or fail the course ($k = 3$ possible outcomes).
- Suppose a test is carried out on n patients, which can be positive, negative and inconclusive ($k = 3$ possible outcomes).
- In a study of n patients the ABO blood group is determined, which can be O, A, B or AB ($k = 4$ possible outcomes).
- The number of votes for party 1, party 2,$\cdots$, party k in an election in which n votes were cast.
- A colour wheel with k different segments is spun n times and we count how often we end up with each colour.

Example 15

Multinomial vs binomial

Rolling a dice

Suppose you roll a dice 10 times and record what number you roll each time. There are 6 possible outcomes $\{1, 2, 3, 4, 5, 6\}$, so this would be a **multinomial** experiment. If you rolled the dice 10 times and recorded how many times you rolled a 3, that would be a **binomial** experiment ($\{3\}$ = success, $\{1, 2, 4, 5, 6\}$ = failure).

In the same way that a binomial experiment will have a binomial distribution, a multinomial experiment will have a multinomial distribution.

Definition 9

Multinomial distribution

Suppose we carry out n independent trials with k possible outcomes. Each time, the probability of observing the j-th outcome is θ_j (with $\theta_1 + \cdots + \theta_k = 1$). Denote by X_j the number of times we observe the j-th outcome. We then say that $\mathbf{X} = (X_1, \cdots, X_k)$ has a multinomial distribution, denoted by

$$\mathbf{X} = (X_1, \cdots X_k) \sim \mathrm{Mu}(n, (\theta_1, \ldots, \theta_k)).$$

Note: The binomial distribution is a special case of the multinomial distribution with $k = 2$. If $X \sim \mathrm{Bi}(n, p)$, then $(X, n - X) \sim \mathrm{Mu}(n, (p, 1 - p))$.

Let's use the following example to state some important properties of the multinomial distribution.

Example 16

Multinomial distribution

Exam scores

10 students take a course where they can either pass the exam on the first sitting, pass the resit exam or fail. If a student fails the first exam, they have to take the resit and if they fail the resit then they will fail the course. We observe how many students fall into each outcome.

Here, $n = 10$, $k = 3$ and let

$X_1 = \{\text{number of students who pass first exam}\}$,

$X_2 = \{\text{number of students who pass resit}\}$,

$X_3 = \{\text{number of students who fail}\}$.

The joint range space for this example consists of all triplets (x_1, x_2, x_3) of non-negative integers that add up to 10, i.e.

$$R_X = \{(10,0,0), (9,1,0), (9,0,1), (8,2,0), (8,0,2), \ldots, (2,0,8), (0,2,8), (1,0,9), (0,1,9), (0,0,10)\}.$$

More generally, X has joint range space

$$R_{\mathbf{X}} = \{(x_1, \ldots, x_k) : x_1, \cdots, x_k = 0, 1, \cdots, n; x_1 + \cdots + x_k = n\}.$$

Proposition 3

Multinomial probability mass function

Suppose that $\mathbf{X} = (X_1, \ldots, X_k) \sim \text{Mu}(n, (\theta_1, \cdots, \theta_k))$. The probability mass function of $\mathbf{X}$, for $x_j \in \mathbb{N}_0$ such that $x_1 + \cdots + x_k = n$, is given by

$$p_{X_1 \cdots X_k}(x_1, \cdots, x_k) = \frac{n!}{x_1! \cdots x_k!} \theta_1^{x_1} \cdots \theta_k^{x_k}.$$

Note: $\mathbb{N}_0$ is the set of natural numbers (including 0), $\mathbb{N}_0 = \{0, 1, 2, \dots\}$.

Example 17

Exam scores

In Example Example 16, we are told an appropriate multinomial distribution is

$$\mathbf{X} = (X_1, X_2, X_3) \sim \text{Mu}(10, (0.6, 0.3, 0.1)).$$

i.e. *on average* 60% of students pass the first exam, 30% pass the resit, and 10% fail. Use this to calculate the probability that

(a) all 10 students pass the first exam. (b) 5 students fail the first exam but all of them pass the resit.

Answer:

We have

$$p_{\mathbf{X}}(x_1, x_2, x_3) = \frac{10!}{x_1!x_2!x_3!}\left(\frac{6}{10}\right)^{x_1}\left(\frac{3}{10}\right)^{x_2}\left(\frac{1}{10}\right)^{x_3}.$$

So

(a)

$$p(\mathbf{X} = [10, 0, 0]) = \frac{10!}{10!0!0!}\left(\frac{6}{10}\right)^{10}\left(\frac{3}{10}\right)^{0}\left(\frac{1}{10}\right)^{0} \approx 0.0060.$$

(b)

$$p(\mathbf{X} = [5, 5, 0]) = \frac{10!}{5!5!0!}\left(\frac{6}{10}\right)^{5}\left(\frac{3}{10}\right)^{5}\left(\frac{1}{10}\right)^{0} \approx 0.0476.$$

Note: These probabilities can be calculated easily in `R` using the `dmultinom` function as follows:

```
dmultinom(c(10,0,0), prob=c(6/10,3/10,1/10))
```

R Console

```
[1] 0.006046618
```

```
dmultinom(c(5,5,0), prob=c(6/10,3/10,1/10))
```

```
R Console
[1] 0.04761711
```

This video provides a detailed explanation of the multinomial probability mass function along with an interesting real-world application.

Video

The multinomial distribution

Duration 6:46

The marginal distributions of each individual random variable, X_j, when viewed individually, are all binomial distributions, with probability of success θ_j, i.e.

$$X_j \sim \text{Bi}(n, \theta_j).$$

Since each individual random variable, X_j, when viewed individually, has a binomial distribution, the expectation and variance are the same as for a binomial distribution (as discussed in Week 3).

The covariance between two of the random variables, X_{j_1} and X_{j_2} (for $j_1 \neq j_2$), is negative,

$$\text{Cov}(X_{j1}, X_{j2}) = -n\theta_{j1}\theta_{j2}$$

This makes intuitive sense. Since all the X_j's must sum to n, if X_{j_1} is larger, it is likely that X_{j_2} will be smaller.

These moments can also be found on the probability formulae sheet which is available on Moodle.

Example 18

Exam scores

In Example 16, an appropriate multinomial distribution is

$$\mathbf{X} = (X_1, X_2, X_3) \sim \text{Mu}(10, (0.6, 0.3, 0.1)).$$

(a) Calculate the probability that at least 7 people pass the first exam and find the expected number and variance of the number of people who will pass the first exam, $E(X_1)$ and $\text{Var}(X_1)$. (b) Calculate the probability that 3 or less people fail the course and find the expected number and variance of the number of people who will fail the course, $E(X_3)$ and $\text{Var}(X_3)$.

Answer:

To calculate these quantities we first need to find the distribution of each random variables. Here we have $X_1 \sim \text{Bi}(10, 0.6)$, $X_2 \sim \text{Bi}(10, 0.3)$, $X_3 \sim \text{Bi}(10, 0.1)$.

(a) $X_1 \sim \text{Bi}(10, 0.6)$,

The probability that at least 7 people pass the first exam is

$$\begin{aligned} P(X_1 \geq 7) &= P(X_1 = 7) + P(X_1 = 8) + P(X_1 = 9) + P(X_1 = 10), \\ &= 0.215 + 0.121 + 0.040 + 0.006, \\ &= 0.382. \end{aligned}$$

$E(X_1) = n\theta_1 = 10 \cdot 0.6 = 6.$

$\text{Var}(X_1) = n\theta_1(1 - \theta_1) = 10 \cdot 0.6 \cdot 0.4 = 2.4.$

Note: These probabilities can be calculated easily in `R` using the `dbinom` function as follows:

```
dbinom(7, 10, prob=6/10) + dbinom(8, 10, prob=6/10) +
  dbinom(9, 10, prob=6/10) + dbinom(10, 10, prob=6/10)
```

R Console

```
[1] 0.3822806
```

or using the `pbinom` function

```
1 - pbinom(6,10, prob=6/10)
```

R Console

```
[1] 0.3822806
```

(b) $X_3 \sim \text{Bi}(10, 0.1)$. The probability that at 3 or less people fail the course is

$$
\begin{aligned}
P(X_3 \leq 3) &= P(X_3 = 3) + P(X_3 = 2) + P(X_3 = 1) + P(X_3 = 0), \\
&= 0.057 + 0.194 + 0.387 + 0.349, \\
&= 0.987.
\end{aligned}
$$

$E(X_3) = n\theta_3 = 10 \cdot 0.1 = 1.$

$\text{Var}(X_3) = n\theta_3(1 - \theta_3) = 10 \cdot 0.1 \cdot 0.9 = 0.9.$

Note: These probabilities can be calculated easily in `R` using the `dbinom` function as follows:

```
dbinom(3, 10, prob=1/10) + dbinom(2, 10, prob=1/10) +
dbinom(1, 10, prob=1/10) + dbinom(0, 10, prob=1/10)
```

R Console

```
[1] 0.9872048
```

or using the `pbinom` function

```
pbinom(3,10, prob=1/10)
```

R Console

```
[1] 0.9872048
```

Task 6

Suppose $\mathbf{X} \sim \text{Mult}(10, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, what is the probability that $\mathbf{X} = (10, 0, 0)$? Or $(5, 2, 4)$?

Task 7

The discrete random vector $\mathbf{X} = (X_1, X_2, X_3) \sim \text{Mult}\left(3, \frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right)$.

a) Write down the joint range space of $\mathbf{X}$.

b) Draw up a table of values of the joint probability mass function, $p_{\mathbf{X}}(x_1, x_2, x_3)$, on this range space.

Learning outcomes for week 4

By the end of week 4, you should be able to:

- use joint probability mass functions to calculate the probability of events relating to bivariate discrete random variables;
- use the joint probability mass function to derive marginal distributions for bivariate discrete random variables;
- use the joint probability mass function to derive conditional probability mass functions for bivariate discrete random variables;
- calculate the covariance and correlation between two discrete random variables;
- calculate if two discrete random variables are independent;
- state and use properties of the multinomial distribution.

A summary of the most important concepts, selected video solutions and written answers to all tasks are provided overleaf.

Week 4 summary

Bivariate discrete random variables

Joint probability mass function (p.m.f.)

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

Key property of the p.m.f.:

$\sum_{(x,y)\in R_{XY}} p_{XY}(x,y) = 1$

Marginal distributions

$$p_X(x) = \sum_{y:(x,y)\in R_{XY}} = p_{XY}(x,y)$$

$$p_Y(y) = \sum_{x:(x,y)\in R_{XY}} = p_{XY}(x,y)$$

Conditional probability mass functions

$$P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{p_{XY}(x,y)}{p_X(x)}$$
$$\text{for } x \in R_X \text{ and } y \text{ such that } (x,y) \in R_{XY}$$

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p_{XY}(x,y)}{p_Y(y)}$$
$$\text{for } y \in R_Y \text{ and } x \text{ such that } (x,y) \in R_{XY}$$

Rearranging gives:

$$p_{XY}(x,y) = P(X = x|Y = y)p_Y(y) = P(Y = y|X = x)p_X(x)$$

Covariance and correlation

The **covariance** between two random variables X and Y is

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

- A **positive** value of the covariance indicates that there is a **positive linear relationship** between two random variables; i.e., higher values of one tend to be observed along with higher values of the other.
- A **negative** value of the covariance indicates that there is a **negative linear relationship** between two random variables; i.e., higher values of one tend to be observed along with lower values of the other.

The **correlation** of X and Y is

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

The sign of the correlation has the same interpretation as the sign of the covariance. Positive correlation indicates a positive linear relationship and negative correlation indicates a negative linear relationship. However, now the value of the correlation informs us how strong the linear relationship between X and Y is.

- If $\rho = 0$ there is no linear relationship. In this case we say that X and Y are *uncorrelated*.
- If $\rho = \pm 1$ there is a perfect linear relationship.

Independence

Two discrete random variables X and Y are called independent if all events relating to X are independent of all events relating to Y.

Let X and Y be two discrete random variables, then the following four statements are equivalent.

(i) X and Y are independent. (ii) The joint probability mass function of X and Y is the product of the marginal probability mass functions, i.e.

$$p_{XY}(x,y) = p_X(x)p_Y(y) \text{ for all } x \text{ and } y.$$

(iii) The conditional distribution of X given $Y = y$ does not depend on y, i.e.

$$p_{X|Y}(x|y) = p_X(x) \quad \text{for all } x \text{ and } y.$$

(iv) The conditional distribution of Y given $X = x$ does not depend on x, i.e.

$$p_{Y|X}(y|x) = p_Y(y) \quad \text{for all } x \text{ and } y.$$

The multinomial distribution

Suppose that $\mathbf{X} = (X_1, \cdots, X_k)$ has a **multinomial distribution**,
$\mathbf{X} = (X_1, \cdots, X_k) \sim \mathrm{Mu}(n, (\theta_1, \cdots, \theta_k))$.

Constraints: $X_1 + \cdots + X_k = n, \quad \theta_1 + \cdots + \theta_k = 1.$

Model

We carry out n independent trials with k possible outcomes. Each time, the probability of observing the j-th outcome is p_j. Denote by X_j the number of times we observe the j-th outcome.

Relationship to binomial

The binomial distribution is a special case of the multinomial distribution with $k = 2$. If $X \sim \mathrm{Bi}(n, \theta)$, then $(X, n - X) \sim \mathrm{Mu}(n, (\theta, 1 - \theta))$.

Probability mass function

Suppose that $\mathbf{X} = (X_1, \ldots, X_k) \sim \mathrm{Mu}(n, (\theta_1, \cdots, \theta_k))$. The probability mass function of $\mathbf{X}$ is, for $x_j \in \mathbb{N}_0$ such that $x_1 + \cdots + x_k = n$, is given by

$$p_{X_1 \cdots X_k}(x_1, \cdots, x_k) = \frac{n!}{x_1! \cdots x_k!} \theta_1^{x_1} \cdots \theta_k^{x_k}.$$

Marginal distribution of X_i

If $\mathbf{X} = (X_1, \ldots, X_k) \sim \mathrm{Mu}(n, (\theta_1, \cdots, \theta_k))$ then,

$$X_j \sim \mathrm{Bi}(n, \theta_j),$$

and

$$E(X_j) = n\theta_j, \quad \mathrm{Var}(X_j) = n\theta_j(1 - \theta_j), \quad \mathrm{Cov}(X_{j1}, X_{j2}) = -n\theta_{j1}\theta_{j2} \quad (\text{for } j_i \neq j_2).$$

Answer 1

a)

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathrm{E}[(X - \mu_X)(Y - \mu_Y)] \\
&= \mathrm{E}(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\
&= \mathrm{E}(XY) - \mathrm{E}(\mu_X Y) - \mathrm{E}(\mu_Y X) + \mathrm{E}(\mu_X \mu_Y) \\
&= \mathrm{E}(XY) - \mu_X \mathrm{E}(Y) - \mu_Y \mathrm{E}(X) + \mu_X \mu_Y \mathrm{E}(1) \\
&= \mathrm{E}(XY) - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y \\
&= \mathrm{E}(XY) - \mu_X \mu_Y \\
&= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).
\end{aligned}$$

b)

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

c)

$$\begin{aligned}\operatorname{Cov}(X,X) &= \operatorname{E}[(X-\mu_X)(X-\mu_X)] \\ &= \operatorname{E}[(X-\mu_X)^2] \\ &= \operatorname{Var}(X).\end{aligned}$$

Answer 2

Distance vision:

1.

$p_{XY}(x,y)$			x		
		1	2	3	$p_Y(y)$
	1	0.2	0.03	0.02	0.25
y	2	0.04	0.3	0.06	0.40
	3	0.01	0.07	0.27	0.35
	$p_X(x)$	0.25	0.40	0.35	1

2.

(a) $P(X=1,Y=3) = p_{XY}(1,3) = 0.01$

(b)

$$\begin{aligned}P(X=1,Y=3) + P(X=3,Y=1) &= p_{XY}(1,3) + p_{XY}(3,1) \\ &= 0.01 + 0.02 = 0.03\end{aligned}$$

(c)

$$
\begin{aligned}
P(X = Y) &= P(X = 1, Y = 1) + P(X = 2, Y = 2) + P(X = 3, Y = 3) \\
&= p_{XY}(1,1) + p_{XY}(2,2) + p_{XY}(3,3) \\
&= 0.20 + 0.30 + 0.27 \\
&= 0.77
\end{aligned}
$$

(d)

$$
\begin{aligned}
P(X < Y) &= P(X = 1, Y = 2) + P(X = 1, Y = 3) + P(X = 2, Y = 3) \\
&= p_{XY}(1,2) + p_{XY}(1,3) + p_{XY}(2,3) \\
&= 0.04 + 0.01 + 0.07 \\
&= 0.12
\end{aligned}
$$

3.

$$E(X) = E(Y) = 1 \cdot 0.25 + 2 \cdot 0.40 + 3 \cdot 0.35 = 2.10$$

$$E(X^2) = E(Y^2) = 1^2 \cdot 0.25 + 2^2 \cdot 0.40 + 3^2 \cdot 0.35 = 5.00$$

$$\text{Var}(X) = \text{Var}(Y) = 5.00 - [2.10]^2 = 5.00 - 4.41 = 0.59$$

4.

$$
\begin{aligned}
E(XY) &= \sum_{(x,y) \in R_{XY}} \sum xy \cdot p_{XY}(x,y) \\
&= (1 \times 1 \times 0.20) + (2 \times 1 \times 0.03) + (3 \times 1 \times 0.02) \\
&+ (1 \times 2 \times 0.04) + (2 \times 2 \times 0.30) + (3 \times 2 \times 0.06) \\
&+ (1 \times 3 \times 0.01) + (2 \times 3 \times 0.07) + (3 \times 3 \times 0.27) \\
&= 0.20 + 0.06 + 0.06 + 0.08 + 1.2 + 0.36 + 0.03 + 0.42 + 2.43 \\
&= 4.84
\end{aligned}
$$

$$
\begin{aligned}
\text{Cov}(X,Y) &= E(XY) - E(X)E(Y) \\
&= 4.84 - 2.10 \cdot 2.10 \\
&= 4.84 - 4.41 \\
&= 0.43
\end{aligned}
$$

$$
\begin{aligned}
\rho(X,Y) &= \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \\
&= \frac{0.43}{\sqrt{0.59 \times 0.59}} \\
&= \frac{0.43}{0.59} \\
&= 0.729
\end{aligned}
$$

Note: In this example, because of the symmetry of $p(x, y)$, Y has the same marginal distribution and therefore expectation and variance as X.

Here is a video worked solution.

Video

Week 4 - Task 2

Duration 17:17

Answer 3

Distance vision continued:

$$P_{X|1}(1|1) = \frac{p_{XY}(1,1)}{p_Y(1)} = \frac{0.20}{0.25} = 0.80$$
$$P_{X|1}(2|1) = \frac{p_{XY}(2,1)}{p_Y(1)} = \frac{0.03}{0.25} = 0.12$$
$$P_{X|1}(3|1) = \frac{p_{XY}(3,1)}{p_Y(1)} = \frac{0.02}{0.25} = 0.08$$

Here is a video worked solution.

Video

Week 4 - Task 3

Duration 2:17

Answer 4

Distance vision continued:

The joint range space is the Cartesian product of the marginal range spaces, but

$$p_{XY}(1,1) = 0.20 \neq 0.0625 = 0.25 \cdot 0.25 = p_X(1) \cdot p_Y(1),$$

thus X and Y are not independent.

Here is a video worked solution.

Video

Week 4 - Task 4

Duration 3:09

Answer 5

X and Y are independent, thus $p_{XY}(x,y) = p_X(x)p_Y(y)$.

$$\begin{aligned} E(XY) &= \sum_{(x,y)\in R_{XY}} xyp_{XY}(x,y) \\ &= \sum_{(x,y)\in R_{XY}} xyp_X(x)p_Y(y) \\ &= \underbrace{\sum_{x\in R_X} xp_X(x)}_{=E(X)} \underbrace{\sum_{yinR_Y} yp_Y(y)}_{=E(Y)} \\ &= E(X)E(Y) \end{aligned}$$

Hence,

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = 0,$$

and thus also $\rho(X,Y) = 0$.

Answer 6

$$P(\mathbf{X} = [10, 0, 0]) = \frac{10!}{10!0!0!}\left(\frac{1}{3}\right)^{10}\left(\frac{1}{3}\right)^{0}\left(\frac{1}{3}\right)^{0} = \left(\frac{1}{3}\right)^{10} \approx 1.69 \times 10^{-5}.$$

$$P(\mathbf{X} = [5, 2, 4]) = 0,$$

since the vector is outside the range space: its elements sum to 11, not 10.

Here is a video worked solution.

Video

Week 4 - Task 6

Duration 4:22

Answer 7

a) The joint range space consists of all triplets (x_1, x_2, x_3) of non-negative integers that add up to 3:

$$R_{\mathbf{X}} = \{(0,0,3), (0,1,2), (0,2,1), (0,3,0), (1,0,2), (1,1,1), (1,2,0), (2,0,1), (2,1,0), (3,0,0)\}.$$

b)

$$p_{\mathbf{X}}(x_1, x_2, x_3) = \frac{3!}{x_1!x_2!x_3!}\left(\frac{1}{2}\right)^{x_1}\left(\frac{1}{3}\right)^{x_2}\left(\frac{1}{6}\right)^{x_3}$$

Here is a video worked solution.

Video

Week 4 - Task 7

Duration 7:13