

# Introduction to Probability & Statistics

Assignment 4, 2024/25

## Practice Questions

**PQ1.** Let  $X$  be a random variable with  $\mathbb{E}[X] = 5$ . What is the expectation of  $3X + 5$ ? If furthermore  $\mathbb{E}[X^2] = 30$ , what is the variance of  $X$ ?

Answer

We can use the linearity of expectation to find that  $\mathbb{E}[3X + 5] = 3\mathbb{E}[X] + 5 = 20$ . The variance is  $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 30 - 5^2 = 5$ .

**PQ2.** I arrive at the train station at 12.00 exactly. My train departs at a time which follows a (continuous) uniform distribution on the interval  $[11.55, 12.15]$ . What is the probability that I miss my train?

Answer

Let  $X$  denote the random time after 11.55 at which the train leaves. The question tells us that  $X \sim \text{Uniform}[0, 20]$ . I miss the train if  $X < 5$ , which has probability

$$\mathbb{P}(X < 5) = \int_0^5 \frac{1}{20} dx = \frac{1}{4}.$$

**PQ3.** Suppose that you have a lecture at 14.00, and that the time taken to travel from your room to the lecture theatre is normally distributed with mean 30 minutes and standard deviation 4 minutes. What is the latest time you should leave your room if you want to be 99% certain that you will not miss the start of the lecture? (Hint: if  $Z \sim N(0, 1)$  then the R function `qnorm(p)` returns the value  $z \in \mathbb{R}$  such that  $\mathbb{P}(Z \leq z) = p$ .)

Answer

Let  $X$  denote the travel time to the lecture:  $X \sim N(30, 16)$ . We wish to find  $x$  such that  $\mathbb{P}(X \leq x) = 0.99$ . Now,

$$\mathbb{P}(X \leq x) = \mathbb{P}\left(\frac{X - 30}{4} \leq \frac{x - 30}{4}\right) = \mathbb{P}\left(Z \leq \frac{x - 30}{4}\right)$$

where  $Z \sim N(0, 1)$ .

We can get hold of this value of  $x$  by using R (or by consulting statistical tables):

`qnorm(0.99)` gives the value 2.326, meaning that  $\mathbb{P}(Z \leq 2.326) = 0.99$ . Thus we require  $(x - 30)/4 = 2.326 \iff x = 39.3$ . Thus the latest you should leave your room is 39.3 minutes before the start of the lecture: i.e. at 13:20.

**PQ4.** A random variable  $Z$  has probability density function

$$f_Z(x) = \begin{cases} \frac{6}{5675}(5x^2 + 3x + 11) & \text{for } 3 \leq x \leq 8 \\ 0 & \text{otherwise.} \end{cases}$$

Would you expect  $\mathbb{E}[Z]$  to lie closer to 3 or to 8? Calculate  $\mathbb{E}[Z]$  and check whether your intuition was correct.

Answer

Since  $f_Z$  is increasing on the interval  $[3, 8]$  we know from the interpretation of expectation as centre of mass that the expectation should lie closer to 8 than to 3. The computation:

$$\mathbb{E}[Z] = \int_3^8 x f_Z(x) dx = \frac{6}{5675} \int_3^8 (5x^3 + 3x^2 + 11x) dx = \frac{2787}{454} = 6.14.$$

**PQ5.** Give an example of a joint probability table for two discrete random variables  $X$  and  $Y$ , each having only two possible values, so that  $F_{X,Y}(5, 6) = 0.4$ ,  $F_X(5) = 0.5$ ,  $F_Y(6) = 0.6$  and  $\mathbb{E}[X] = 10$ ,  $\mathbb{E}[Y] = 4$ .

Answer

One possible example would be

$y \backslash x$	0	20	$p_Y(y)$
0	0.4	0.2	0.6
10	0.1	0.3	0.4
$p_X(x)$	0.5	0.5	1

**PQ6.** The joint probability mass function  $p_{X,Y}(x, y)$  of two random variables  $X$  and  $Y$  is summarised by the following table:

$x \backslash y$	-1	0	1
4	$\eta - 1/16$	$1/4 - \eta$	0
5	$1/8$	$3/16$	$1/8$
6	$\eta + 1/16$	$1/16$	$1/4 - \eta$

where  $\eta$  is a real number.

- Extend the table by including also the marginal probabilities, i.e., the values of the probability mass functions  $p_X$  and  $p_Y$ .
- Which are the valid choices for  $\eta$ ?
- Is there a value of  $\eta$  for which  $X$  and  $Y$  are independent?

Answer

- We extend the probability table to also include the marginal probability mass functions  $p_X$  and  $p_Y$ :

$x \backslash y$	-1	0	1	$p_X(x)$
4	$\eta - 1/16$	$1/4 - \eta$	0	$3/16$
5	$1/8$	$3/16$	$1/8$	$7/16$
6	$\eta + 1/16$	$1/16$	$1/4 - \eta$	$3/8$
$p_Y(y)$	$2\eta + 1/8$	$1/2 - \eta$	$3/8 - \eta$	1

- b) All entries of the probability table must be non-negative and they must sum up to 1. In order for  $p_{X,Y}(4, -1)$  to be non-negative we need  $\eta \geq 1/16$ . In order for  $p_{X,Y}(4, 0)$  and  $p_{X,Y}(6, 1)$  to be non-negative we need  $\eta \leq 1/4$ . The sum over all entries is not affected by the value of  $\eta$ , so does not give any additional constraints. Therefore any  $\eta \in [1/16, 1/4]$  is a valid choice.
- c) It is easy to find counterexamples to the factorisation of the joint probability mass function that would have to hold if  $X$  and  $Y$  were independent. For example

$$p_X(4)p_Y(1) = \frac{3}{16} \left( \frac{3}{8} - \eta \right) \neq 0 = p_{X,Y}(4, 1)$$

unless  $\eta = 3/8$ . However the value  $\eta = 3/8$  is not allowed, and hence  $X$  and  $Y$  can never be independent.

**PQ7.** Let  $X$  and  $Y$  be random variables. Show that  $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ .

Answer

We start from the definition of covariance, and use linearity of expectation:

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

## Assignment Questions – answers to be uploaded

**AQ1.** Buses leave campus for the train station every 15 minutes, at 0, 15, 30 and 45 minutes past the hour. If a student arrives at the bus stop at a time that follows a (continuous) uniform distribution on the interval between 10.00 and 10.30, find the probability that they wait

- less than 5 minutes for a bus;
- at least 8 minutes for a bus.

Answer

Let  $Y$  denote the number of minutes past 10.00 that the student arrives at the bus stop:  $Y \sim \text{Uniform}[0, 30]$ . [1]

- They will wait less than 5 minutes if and only if  $10 \leq Y \leq 15$  or  $25 \leq Y \leq 30$ . This occurs with probability

$$\mathbb{P}(10 \leq Y \leq 15) + \mathbb{P}(25 \leq Y \leq 30) = \int_{10}^{15} \frac{1}{30} dy + \int_{25}^{30} \frac{1}{30} dy = \frac{1}{3}.$$

[2]

- Similarly, they will wait at least 8 minutes if they arrive between 10.00 and 10.07, or between 10.15 and 10.22. This has probability  $14/30 = 7/15$ . [2]

**AQ2.** Let  $X \sim \text{Geom}(p)$ . Calculate  $\mathbb{E}[h(X)]$ , where  $h(x) = e^{tx}$  for some  $t > 0$ . For what values of  $t$  is  $\mathbb{E}[h(X)] < \infty$ ?

### Answer

We use the formula for the expectation of a function of a discrete random variable:

$$\begin{aligned}\mathbb{E}[h(X)] &= \sum_{k=1}^{\infty} h(k)p(1-p)^{k-1} = \sum_{k=1}^{\infty} e^{tk}p(1-p)^{k-1} \\ &= pe^t \sum_{k=1}^{\infty} [e^t(1-p)]^{k-1} = pe^t \sum_{k=0}^{\infty} [e^t(1-p)]^k \\ &= \frac{pe^t}{1 - e^t(1-p)}.\end{aligned}$$

[4]

This final step requires  $e^t(1-p) < 1$ . (Otherwise the geometric sum does not converge to a finite limit.) [1]

**AQ3.** Show that if  $Z$  is a standard normal random variable then, for  $x > 0$ ,

- a)  $\mathbb{P}(Z > x) = \mathbb{P}(Z < -x)$ ;
- b)  $\mathbb{P}(|Z| > x) = 2\mathbb{P}(Z > x)$ ;
- c)  $\mathbb{P}(|Z| < x) = 2\mathbb{P}(Z < x) - 1$ .

Hint: express the probabilities in terms of integrals over the density function  $\phi$ , and use the fact that  $\phi$  is an even function (i.e.  $\phi(z) = \phi(-z)$ ).

### Answer

There are many ways to show these identities. We use the hint about the symmetry of the density function of a standard normal random variable:

$$\phi(-z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(-z)^2}{2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) = \phi(z).$$

a)

$$\mathbb{P}(Z > x) = \int_x^{\infty} \phi(z)dz = \int_{-\infty}^{-x} \phi(-u)du = \int_{-\infty}^{-x} \phi(u)du = \mathbb{P}(Z < -x);$$

[1]

b)

$$\mathbb{P}(|Z| > x) = \mathbb{P}(Z > x) + \mathbb{P}(Z < -x) = 2\mathbb{P}(Z > x),$$

where the last equality follows from part (a). [2]

c)

$$\begin{aligned}\mathbb{P}(|Z| < x) &= 1 - \mathbb{P}(|Z| > x) = 1 - 2\mathbb{P}(Z > x) \\ &= 1 - 2(1 - \mathbb{P}(Z < x)) = 2\mathbb{P}(Z < x) - 1,\end{aligned}$$

where the second equality follows from part (b). [2]

**AQ4.** Let  $X : \Omega \rightarrow \{1, 2\}$  and  $Y : \Omega \rightarrow \{0, 1\}$  be two discrete random variables. The following is a partial table of their joint and their marginal mass functions:

$y \backslash x$	1	2	$p_Y(y)$
0	1/6	1/2	

$y \backslash x$	1	2	$p_Y(y)$
1			
$p_X(x)$	5/12		1

- Fill in the missing values.
- Determine the joint distribution function of  $X$  and  $Y$ .
- Calculate  $\mathbb{E}[X]$  and  $\mathbb{E}[Y]$ .
- Let  $Z = XY$ . Calculate  $\mathbb{E}[Z]$ .

#### Answer

- The missing entries in the probability table are determined by the requirement that summing the joint probabilities across a row or across a column in the table gives the corresponding marginal probability and by the requirement that the marginal probabilities for  $X$  as well as those for  $Y$  have to add up to 1. So first we determine  $p_Y(0) = 1/6 + 1/2 = 2/3$ . Then we can determine  $p_Y(1) = 1 - p_Y(0) = 1 - 2/3 = 1/3$  and  $p_X(2) = 1 - p_X(1) = 1 - 5/12 = 7/12$ . Finally we determine  $p_{X,Y}(1,1) = p_X(1) - p_{X,Y}(1,0) = 5/12 - 1/6 = 1/4$  and  $p_{X,Y}(2,1) = p_X(2) - p_{X,Y}(2,0) = 7/12 - 1/2 = 1/12$ . [1]

$y \backslash x$	1	2	$p_Y(y)$
0	1/6	1/2	2/3
1	1/4	1/12	1/3
$p_X(x)$	5/12	7/12	1

- b) The joint distribution function  $F_{X,Y}(x,y)$  is by definition given by  $\mathbb{P}(X \leq x, Y \leq y)$ . So for example

$$F_{X,Y}(1.5, 1.5) = p_{X,Y}(1, 0) + p_{X,Y}(1, 1) = \frac{1}{6} + \frac{1}{4} = \frac{5}{12}.$$

By doing more such calculations we find that

$$F_{X,Y} = \begin{cases} 0 & \text{if } x < 1 \text{ or } y < 0 \\ 1/6 & \text{if } x \in [1, 2) \text{ and } y \in [0, 1) \\ 5/12 & \text{if } x \in [1, 2) \text{ and } y \geq 1 \\ 2/3 & \text{if } x \geq 2 \text{ and } y \in [0, 1) \\ 1 & \text{if } x \geq 2 \text{ and } y \geq 1. \end{cases}$$

[2]

- c) For calculating the expectations of  $X$  and  $Y$  we can use their marginal mass functions:

$$\mathbb{E}[X] = 1 \cdot p_X(1) + 2 \cdot p_X(2) = 1 \cdot \frac{5}{12} + 2 \cdot \frac{7}{12} = \frac{19}{12}$$

and

$$\mathbb{E}[Y] = 0 \cdot p_Y(0) + 1 \cdot p_Y(1) = p_Y(1) = \frac{1}{3}.$$

[1]

- d) The random variable  $Z = XY$  can take the possible values 0, 1 and 2 with probabilities

$$p_Z(0) = p_{X,Y}(1, 0) + p_{X,Y}(2, 0) = p_Y(0) = \frac{2}{3}$$

$$p_Z(1) = p_{X,Y}(1, 1) = \frac{1}{4}, \quad p_Z(2) = p_{X,Y}(2, 1) = \frac{1}{12}.$$

Thus

$$\mathbb{E}[Z] = 1 \cdot p_Z(1) + 2 \cdot p_Z(2) = \frac{1}{4} + 2 \cdot \frac{1}{12} = \frac{5}{12}.$$

[1]

## Other Questions (for seminars / extra practice)

**OQ1.** A married couple decide to have children until they have at least one child of each sex: let  $X$  denote the total number of children that they have. The probability of any one child being a boy is  $1/2$  (with the sex of each child being independent of all the others).

- a) What is the mass function of  $X$ ? (I.e. write down  $\mathbb{P}(X = n)$  for all  $n \in \mathbb{N}$ .)  
b) Show that

$$\mathbb{E}[X] = 3.$$

Hint: you may find it useful to refer to the result from lectures that if  $Y \sim \text{Geom}(p)$  then  $\mathbb{E}[Y] = 1/p$ .

### Answer

- a) Clearly the couple need to have at least two children, so  $\mathbb{P}(X = 1) = 0$ . For  $n \geq 2$ , there are two ways in which the couple can have exactly  $n$  children: either they have  $n-1$  boys

in a row, and then a girl; or they have  $n - 1$  girls and then a boy. Each of these possibilities has probability  $(1/2)^n$ . Thus

$$\mathbb{P}(X = n) = (1/2)^n + (1/2)^n = (1/2)^{n-1}, \quad n \geq 2.$$

b) Here are two possible ways of calculating  $\mathbb{E}[X]$ .

**Method 1:** We use the usual formula for the expectation of a discrete random variable:

$$\mathbb{E}[X] = \sum_{n=2}^{\infty} n \mathbb{P}(X = n) = \sum_{n=2}^{\infty} n (1/2)^{n-1}$$

Using the hint, we know that if  $Y \sim \text{Geom}(1/2)$  then  $\mathbb{E}[Y] = 2$ . That is,

$$\sum_{n=1}^{\infty} n (1/2) (1/2)^{n-1} = 2.$$

We now manipulate our expression for the expectation, until it looks like something involving this result:

$$\begin{aligned} \mathbb{E}[X] &= \sum_{n=2}^{\infty} n (1/2)^{n-1} = 2 \sum_{n=2}^{\infty} n (1/2) (1/2)^{n-1} \\ &= 2 \left[ \sum_{n=1}^{\infty} n (1/2) (1/2)^{n-1} - 1/2 \right] \\ &= 2[2 - 1/2] = 3. \end{aligned}$$

**Method 2:** If we let  $Y = X - 1$  then this random variable takes values in the set  $\mathbb{N}$  and has mass function

$$\mathbb{P}(Y = n) = \mathbb{P}(X - 1 = n) = \mathbb{P}(X = n + 1) = (1/2)^n$$

for  $n \in \mathbb{N}$ . Thus  $Y \sim \text{Geom}(1/2)$ . It follows that  $\mathbb{E}[X] = \mathbb{E}[Y + 1] = \mathbb{E}[Y] + 1 = 2 + 1 = 3$ . (Here we're effectively observing that the couple start by having one child, who could be of either sex; they then need to have an additional random number of children until they have one of the opposite sex to the first – this is like repeating independent Bernoulli trials, with “success” meaning that they have a child of the opposite sex.)

**OQ2.** Let  $X$  be a discrete random variable. Show that for all functions  $h_1, h_2 : \mathbb{R} \rightarrow \mathbb{R}$ ,

$$\mathbb{E}[h_1(X) + h_2(X)] = \mathbb{E}[h_1(X)] + \mathbb{E}[h_2(X)].$$

**Answer**

Let  $h(x) = h_1(x) + h_2(x)$ . From the formula for the expectation of a function of a discrete random variable it follows that

$$\begin{aligned} \mathbb{E}[h(X)] &= \sum_{k \in X(\Omega)} h(k) p_X(k) \\ &= \sum_{k \in X(\Omega)} (h_1(k) + h_2(k)) p_X(k) \\ &= \sum_{k \in X(\Omega)} h_1(k) p_X(k) + \sum_{k \in X(\Omega)} h_2(k) p_X(k) \\ &= \mathbb{E}[h_1(X)] + \mathbb{E}[h_2(X)]. \end{aligned}$$

**OQ3.** Let  $X$  and  $Y$  be random variables and let  $r, s, t, u \in \mathbb{R}$ . Show that

$$\rho(rX + s, tY + u) = \begin{cases} \rho(X, Y) & \text{if } rt > 0 \\ 0 & \text{if } rt = 0 \\ -\rho(X, Y) & \text{if } rt < 0 \end{cases}$$

where  $\rho(X, Y)$  denotes the correlation coefficient of  $X$  and  $Y$ .

Answer

Let us first assume that  $\text{Var}(X)\text{Var}(Y) > 0$  and  $rt > 0$ . Then the definition of the correlation coefficient gives

$$\rho(rX + s, tY + u) = \frac{\text{Cov}(rX + s, tY + u)}{\sqrt{\text{Var}(rX + s)\text{Var}(tY + u)}}. \quad (1)$$

We already know that

$$\text{Var}(rX + s) = r^2\text{Var}(X), \quad \text{Var}(tY + u) = t^2\text{Var}(Y). \quad (2)$$

We need to derive a similar transformation rule for the covariance.

$$\begin{aligned} \text{Cov}(rX + s, tY + u) &= \mathbb{E}[(rX + s - \mathbb{E}[rX + s])(tY + u - \mathbb{E}[tY + u])] \\ &= \mathbb{E}[(rX + s - (r\mathbb{E}[X] + s))(tY + u - (t\mathbb{E}[Y] + u))] \\ &= \mathbb{E}[r(X - \mathbb{E}[X])t(Y - \mathbb{E}[Y])] \\ &= rt\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= rt\text{Cov}(X, Y), \end{aligned} \quad (3)$$

where we repeatedly used the linearity of expectation. Using the transformation rules Equation 2 and Equation 3 in Equation 1 gives

$$\rho(rX + s, tY + u) = \frac{rt}{\sqrt{r^2t^2}} \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

The statement now follows from the observation that

$$\frac{rt}{\sqrt{r^2t^2}} = \begin{cases} 1 & \text{if } rt > 0 \\ -1 & \text{if } rt < 0. \end{cases}$$

In case  $\text{Var}(X)\text{Var}(Y) = 0$  or  $rt = 0$  also  $\text{Var}(rX + s)\text{Var}(tY + u) = rt\text{Var}(X)\text{Var}(Y) = 0$ , and thus  $\rho(rX + s, tY + u) = 0$  by definition. This agrees with the statement because when  $\text{Var}(X)\text{Var}(Y) = 0$  also  $\rho(X, Y) = 0$ .

**OQ4.** Let  $X \sim \text{Uniform}(0, a)$  for some  $a > 0$ . Show that for any  $n \in \mathbb{N}$ ,

$$\mathbb{E}[X^n] = \frac{a^n}{n+1}.$$

Use this to determine  $\rho(X, X^2)$ , and show that this does not depend upon the value of  $a$ .

Answer

For  $n \in \mathbb{N}$  we calculate

$$\mathbb{E}[X^n] = \int_{-\infty}^{\infty} x^n f_X(x) dx = \int_0^a \frac{x^n}{a} dx = \frac{1}{a} \left[ \frac{x^{n+1}}{n+1} \right]_0^a = \frac{a^n}{n+1}.$$



Now we calculate the covariance of  $X$  and  $X^2$ :

$$\text{Cov}(X, X^2) = \mathbb{E}[X^3] - \mathbb{E}[X] \mathbb{E}[X^2] = \frac{a^3}{4} - \frac{a^2}{3} \frac{a}{2} = \frac{a^3}{12}.$$

We also have

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{a^2}{3} - \left(\frac{a}{2}\right)^2 = \frac{a^2}{12}$$

and

$$\text{Var}(X^2) = \mathbb{E}[X^4] - \mathbb{E}[X^2]^2 = \frac{a^4}{5} - \left(\frac{a^2}{3}\right)^2 = \frac{4a^4}{45}.$$

Finally, we calculate

$$\rho(X, X^2) = \frac{\text{Cov}(X, X^2)}{\sqrt{\text{Var}(X) \text{Var}(X^2)}} = \frac{a^3/12}{\sqrt{a^2/12 \cdot 4a^4/45}} = \frac{a^3/12}{\sqrt{a^6/135}} = \frac{\sqrt{135}}{12} = \frac{\sqrt{15}}{4},$$

which doesn't depend upon  $a$ .

**OQ5.** A bag contains 3 cubes, 4 pyramids and 7 spheres. An object is drawn randomly from the bag and its type is recorded. Then the object is replaced. This is repeated 20 times.

- Let  $C_i$  be the indicator random variable for the event that the  $i$ -th draw gives a cube, for  $i = 1, \dots, 20$ . Calculate  $\mathbb{E}[C_i]$ ,  $\mathbb{E}[C_i^2]$  and  $\mathbb{E}[C_i C_j]$  for  $i \neq j$ .
- Let  $C$  be the number of times a cube was drawn, Use that  $C = \sum_{i=1}^{20} C_i$  to calculate  $\mathbb{E}[C]$  and  $\text{Var}(C)$ .
- Let  $S_i$  be the indicator random variable for the event that the  $i$ -th draw gives a sphere. Calculate  $\mathbb{E}[C_i S_i]$  and  $\mathbb{E}[C_i S_j]$  for  $i \neq j$ .
- Let  $S$  be the number of times a sphere was drawn. Use the above results to calculate  $\mathbb{E}[CS]$ ,  $\text{Cov}(C, S)$ ,  $\rho(C, S)$ .

**Answer**

- As three of the 14 shapes are cubes, the probability to draw a cube is  $3/14$ . Hence  $C_i \sim \text{Bern}(3/14)$ . This immediately gives

$$\mathbb{E}[C_i] = \mathbb{E}[C_i^2] = \frac{3}{14}.$$

For  $i \neq j$  the event that the  $i$ -th draw gives a cube and the event that the  $j$ -th cube gives a draw are independent (because we put the shape back after each draw). Thus the indicator random variables  $C_i$  and  $C_j$  for these events are also independent and thus

$$\mathbb{E}[C_i C_j] = \mathbb{E}[C_i] \mathbb{E}[C_j] = \left(\frac{3}{14}\right)^2 = \frac{9}{196}.$$

- The linearity of expectation gives

$$\mathbb{E}[C] = \mathbb{E}\left[\sum_{i=1}^{20} C_i\right] = \sum_{i=1}^{20} \mathbb{E}[C_i] = 20 \frac{3}{14} = \frac{30}{7}.$$

Because the  $C_i$  are independent of each other, the variance of their sum equals the sum of their variances:

$$\text{Var}(C) = \text{Var}\left(\sum_{i=1}^{20} C_i\right) = \sum_{i=1}^{20} \text{Var}(C_i) = 20 \frac{3}{14} \frac{11}{14} = \frac{165}{49}.$$

- c. We observe that  $C_i S_i = 0$  because on the same draw one can not simultaneously have a cube and a sphere. Thus also  $\mathbb{E}[C_i S_i] = 0$ . If  $i \neq j$  we can use independence to factorise the expectation:

$$\mathbb{E}[C_i S_j] = \mathbb{E}[C_i] \mathbb{E}[S_j] = \frac{3}{14} \frac{1}{2} = \frac{3}{28},$$

where we used that the probability of drawing a sphere is  $1/2$ .

- d. We have

$$\mathbb{E}[CS] = \mathbb{E}\left[\sum_{i=1}^{20} C_i \sum_{j=1}^{20} S_j\right] = \sum_{i=1}^{20} \sum_{j=1}^{20} \mathbb{E}[C_i S_j].$$

We split the sum over all pairs  $(i, j)$  into the pairs where  $i \neq j$  and the pairs  $(i, i)$ , so

$$\mathbb{E}[CS] = \sum_{i=1}^{20} \sum_{\substack{j=1 \\ j \neq i}}^{20} \mathbb{E}[C_i S_j] + \sum_{i=1}^{20} \mathbb{E}[C_i S_i].$$

Using our above results for  $\mathbb{E}[C_i S_i]$  and  $\mathbb{E}[C_i S_j]$  and recognising that there are  $20 \cdot 19 = 380$  pairs where  $i \neq j$  this gives us

$$\mathbb{E}[CS] = \sum_{i=1}^{20} \sum_{\substack{j=1 \\ j \neq i}}^{20} \frac{3}{28} + \sum_{i=1}^{20} 0 = 380 \frac{3}{28} = \frac{285}{7}.$$

We also calculate

$$\mathbb{E}[S] = \mathbb{E}\left[\sum_{i=1}^{20} S_i\right] = \sum_{i=1}^{20} \mathbb{E}[S_i] = 20 \frac{1}{2} = 10.$$

The covariance can then be calculated as

$$\text{Cov}(C, S) = \mathbb{E}[CS] - \mathbb{E}[C] \mathbb{E}[S] = \frac{285}{7} - \frac{30}{7} 10 = -\frac{15}{7}.$$

To calculate the correlation coefficient we also need

$$\text{Var}(S) = \text{Var}\left(\sum_{i=1}^{20} S_i\right) = \sum_{i=1}^{20} \text{Var}(S_i) = 20 \frac{1}{2} \frac{1}{2} = 5.$$

The correlation coefficient is

$$\rho(C, S) = \frac{\text{Cov}(C, S)}{\sqrt{\text{Var}(C) \text{Var}(S)}} = -\sqrt{\frac{3}{11}} \approx -0.5222.$$

**OQ6.** Prove that binomial coefficients satisfy the identity

$$n \binom{n-1}{r-1} = r \binom{n}{r}.$$

Use this to find  $\mathbb{E}[X]$  and  $\text{Var}(X)$ , where  $X \sim \text{Bin}(n, p)$ .

Answer

First we prove the identity:

$$n \binom{n-1}{r-1} = n \frac{(n-1)!}{(r-1)!(n-r)!} = r \frac{n!}{r!(n-r)!} = r \binom{n}{r}.$$

For the mean and variance, remember that, since  $p_X(\cdot)$  is a mass function, it must sum to one. That is,

$$\sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = 1. \quad (4)$$

Now,

$$\begin{aligned} \mathbb{E}[X] &= \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^n n \binom{n-1}{k-1} p^k (1-p)^{n-k} \quad (\text{by our identity}) \\ &= np \sum_{k=1}^n \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \\ &= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{(n-1)-j} \quad (\text{putting } j = k-1) \\ &= np, \end{aligned}$$

thanks to Equation 4.

Furthermore,

$$\begin{aligned} \mathbb{E}[X(X-1)] &= \sum_{k=0}^n k(k-1) \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^n k(k-1) \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^n \frac{n!}{(n-k)!(k-2)!} p^k (1-p)^{n-k} \\ &= n(n-1)p^2 \sum_{k=2}^n \frac{(n-2)!}{((n-2)-(k-2))!(k-2)!} p^{k-2} (1-p)^{(n-2)-(k-2)} \\ &= n(n-1)p^2 \sum_{j=0}^{n-2} \binom{n-2}{j} p^j (1-p)^{(n-2)-j} \quad (\text{putting } j = k-2) \\ &= n(n-1)p^2, \end{aligned}$$

again thanks to Equation 4. It follows that

$$\mathbb{E}[X^2] = n(n-1)p^2 + np,$$

and so

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = np(1-p).$$

**OQ7. [Harder]** Consider a random variable  $X \sim \text{Uniform}[a, b]$ , where  $a$  and  $b$  are unknown. You are told that

$$\mathbb{P}(X < 2) = 1/3 \quad \text{and} \quad \mathbb{P}(1 < X \leq 3) = 1/2.$$

Given this information, find  $a$  and  $b$ .

Answer

From the first equation we immediately know that  $a < 2 < b$ . Now, for a continuous random variable, we obtain the probability that it lies in an interval  $(c, d)$  by integrating the density

function over that interval, i.e.

$$\mathbb{P}(c \leq X \leq d) = \int_c^d f_X(x) dx.$$

Since  $X \sim \text{Uniform}[a, b]$ , we know that

$$f_X(x) = \begin{cases} 1/(b-a) & x \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

Thus we obtain

$$1/3 = \mathbb{P}(X < 2) = \mathbb{P}(a \leq X < 2) = \int_a^2 1/(b-a) dx = (2-a)/(b-a). \quad (5)$$

In order to use the second equation ( $\mathbb{P}(1 < X \leq 3) = 1/2$ ) in the same way, we have two possibilities to consider:

1.  $a < 1$
2.  $1 \leq a < 2$

Suppose first that  $a < 1$ . Then

$$1/2 = \mathbb{P}(1 < X \leq 3) = \int_1^3 1/(b-a) dx = 2/(b-a), \quad (6)$$

since the density function  $f_X$  is equal to  $1/(b-a)$  for all  $x \in [1, 3]$  if  $a < 1$ .

If  $1 \leq a$  however, then instead we obtain

$$1/2 = \mathbb{P}(1 < X \leq 3) = \int_1^a 0 dx + \int_a^3 1/(b-a) dx = (3-a)/(b-a). \quad (7)$$

We now have to solve these simultaneous equations in order to find  $a$  and  $b$ . If we assume that  $1 \leq a < 2$ , then we must try to solve Equation 5 and Equation 7 together; but this gives

$$2(3-a) = 3(2-a),$$

resulting in  $a = 0$ . But this contradicts our assumption that  $1 \leq a$ !

So it must be the case that  $a < 1$ : now we must solve Equation 5 and Equation 6, and this *is* possible, with  $a = 2/3$  and  $b = 14/3$ .

**OQ8. [Harder]** Let  $X$  and  $Y$  be two independent geometrically distributed random variables with parameter  $p$ , i.e.,  $X \sim \text{Geom}(p)$  and  $Y \sim \text{Geom}(p)$ . For any natural numbers  $i$  and  $n$  with  $i < n$  calculate the conditional probability  $\mathbb{P}(X = i | X + Y = n)$ . Describe in words the meaning in terms of Bernoulli trials of what you just calculated.

Answer

According to the definition of conditional probability,

$$\mathbb{P}(X = i | X + Y = n) = \frac{\mathbb{P}(X = i, X + Y = n)}{\mathbb{P}(X + Y = n)}.$$

For the numerator we can use that the event  $\{X = i, X + Y = n\}$  is the event  $\{X = i, Y = n - i\}$ . We then know that the independence of  $X$  and  $Y$  implies the factorisation of that probability:

$$\mathbb{P}(X = i, X + Y = n) = \mathbb{P}(X = i, Y = n - i) = \mathbb{P}(X = i) \mathbb{P}(Y = n - i).$$

We can now substitute in the probability mass function for the geometric distribution with parameter  $p$ :

$$\mathbb{P}(X = i) = (1 - p)^{i-1}p$$

and thus

$$\mathbb{P}(Y = n - i) = (1 - p)^{n-i-1}p.$$

This gives

$$\mathbb{P}(X = i, X + Y = n) = (1 - p)^{i-1}p(1 - p)^{n-i-1}p = (1 - p)^{n-2}p^2.$$

Note that this is independent of  $i$ .

For the denominator we use the partition theorem to write

$$\mathbb{P}(X + Y = n) = \sum_{i=1}^{n-1} \mathbb{P}(X = i, X + Y = n).$$

From our calculation above we see that every term in the sum is the same, so

$$\mathbb{P}(X + Y = n) = (n - 1)\mathbb{P}(X = i, X + Y = n).$$

Putting this all together we finally find that

$$\mathbb{P}(X = i | X + Y = n) = \frac{\mathbb{P}(X = i, X + Y = n)}{\mathbb{P}(X + Y = n)} = \frac{1}{n - 1}.$$

A geometric random variable counts the number of turns until the first success in repeated Bernoulli trials. Therefore the sum  $X + Y$  of two identical and independent geometric random variables counts the number of turns until the *second* success. So the conditional probability we calculated is the probability that the first success happens on a particular trial  $i$  given that the second success happens on the  $n$ -th trial. The result shows that the first success is then equally likely to occur on any of the  $n - 1$  trials before the  $n$ -th trial.