Section 4 – Solutions

Review

- **Random Variable (rv)**: A numeric function $X : \Omega \to \mathbb{R}$ of the outcome.
- **Range/Support:** The support/range of a random variable X, denoted Ω_X, is the set of all possible values that X can take on.
- **Discrete Random Variable (drv)**: A random variable taking on a countable (either finite or countably infinite) number of possible values.
- **Probability Mass Function (pmf) for a discrete random variable X**: a function $p_X : \Omega_X \to [0, 1]$ with $p_X(x) = \mathbb{P}(X = x)$ that maps possible values of a discrete random variable to the probability of that value happening, such that $\sum_x p_X(x) = 1$.
- **Cumulative Distribution Function (CDF) for a random variable X**: a function $F_X : R \to R$ with $F_X(x) = \mathbb{P}(X \leqslant x)$
- **Expectation (expected value, mean, or average)**: The expectation of a discrete random variable is defined to be $\mathbb{E}[X] = \sum_x x p_X(x) = \sum_x x\mathbb{P}(X = x)$. The expectation of a function of a discrete random variable $g(X)$ is $\mathbb{E}[g(X)] = \sum_x g(x)p_X(x)$.
- **Linearity of Expectation**: Let X and Y be random variables, and $a, b, c \in \mathbb{R}$. Then, $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$. Also, for any random variables $X_1, \ldots, X_n$,

$$\mathbb{E}[X_1 + X_2 + \ldots + X_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \ldots + \mathbb{E}[X_n].$$

- **Variance**: Let X be a random variable and $\mu = \mathbb{E}[X]$. The variance of X is defined to be $\mathrm{Var}(X) = \mathbb{E}[(X - \mu)^2]$. Notice that since this is an expectation of a non-negative random variable ($(X - \mu)^2$), variance is always non-negative. With some algebra, we can simplify this to $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.
- **Standard Deviation**: Let X be a random variable. We define the standard deviation of X to be the square root of the variance, and denote it $\sigma = \sqrt{\mathrm{Var}(X)}$.
- **Property of Variance**: Let $a, b \in \mathbb{R}$ and let X be a random variable. Then, $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.
- **Independence**: Random variables X and Y are independent iff

$$\forall x \forall y, \quad \mathbb{P}(X = x \cap Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$

  In this case, we have $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ (the converse is not necessarily true).
- **i.i.d. (independent and identically distributed)**: Random variables $X_1, \ldots, X_n$ are i.i.d. (or iid) iff they are independent and have the same probability mass function.
- **Variance of Independent Variables**: If X is independent of Y, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. This depends on independence, whereas linearity of expectation always holds. Note that this combined with the above shows that $\forall a, b, c \in \mathbb{R}$ and if X is independent of Y, $\mathrm{Var}(aX + bY + c) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)$.

Task 1 – Content Review

a) True or false: the range of a random variable X is the set of probabilities corresponding to the possible values X can take on.

False. The range (or *support*) of a random variable is the set of all possible values it can take on.

b) What is the relationship between standard deviation and variance of a random variable X?

- ◯ $\sigma = (\mathrm{Var}(X))^2$
- ◯ $\sigma = \mathrm{Var}(X^2)$
- ◯ $\mathrm{Var}(X) = \sigma^2$

$\mathrm{Var}(X) = \sigma^2$ (or $\sigma = \sqrt{\mathrm{Var}(X)}$ in the above review)

c) Let X be the random variable representing the outcome of taking the sum of a 3-dice roll of 6-sided dice. Which function would you use to determine the probability that $X = 7$?

- ◯ CDF (cumulative distribution function)
- ◯ PMF (probability mass function)

PMF. We use the PMF when we want to find the probability of a specific value of a random variable occurring.

d) Let X be the random variable representing the outcome of taking the sum of a 3-dice roll of 6-sided dice. Which function would you use to determine the probability that $X \leqslant 7$?

- ◯ CDF (cumulative distribution function)
- ◯ PMF (probability mass function)

CDF. The CDF gives us exactly $\mathbb{P}(X \leqslant x)$.

e) A random variable X has the PMF

$$p_X(x) = \begin{cases} 1/4 & x = -1 \\ 1/4 & x = 0 \\ 1/2 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

What is $\mathbb{E}[X]$?

- ◯ -1/4
- ◯ 3/4
- ◯ 1
- ◯ 2

3/4.

$$\mathbb{E}[X] = \sum_{x \in \Omega_X} x p_X(x) = -1 \cdot 1/4 + 0 \cdot 1/4 + 2 \cdot 1/2 = 3/4.$$

f) A random variable X has the PMF

$$p_X(x) = \begin{cases} 1/4 & x = -1 \\ 1/4 & x = 0 \\ 1/2 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

What is $\mathrm{Var}[X]$?

◯ $3/4$
◯ 1
◯ $((1/4)+2)-((\frac{3}{4})^2)=27/16$
◯ $((1/4)+2)+((\frac{3}{4})^2)=45/16$

$27/16$.

$$\begin{aligned}\text{Var}[X]=\mathbb{E}[X^2]-\mathbb{E}[X]^2 &= \sum_{x\in\Omega_X} x^2 p_X(x) - \sum_{x\in\Omega_X} x p_X(x)\\ &= ((-1)^2\cdot 1/4 + 0^2\cdot 1/4 + 2^2\cdot 1/2) - ((3/4)^2)\\ &= 27/16\end{aligned}$$

Task 2 – Identify that range!

Identify the support/range Ω_X of the random variable X, if X is...

a) The sum of two rolls of a six-sided die.

X takes on every integer value between the min sum 2, and the max sum 12.
$\Omega_X = \{2, 3, ..., 12\}$

b) The number of lottery tickets I buy until I win it.

X takes on all positive integer values (I may never win the lottery).
$\Omega_X = \{1, 2, ...\} = \mathbb{N}$

c) The number of heads in n flips of a coin with $0 < \mathbb{P}(\text{head}) < 1$.

X takes on every integer value between the min number of heads 0, and the max n.
$\Omega_X = \{0, 1, ..., n\}$

d) The number of heads in n flips of a coin with $\mathbb{P}(\text{head}) = 1$.

Since $\mathbb{P}(\text{head}) = 1$, we are guaranteed to get n heads in n flips.
$\Omega_X = \{n\}$

Linearity of Expectation Problems
The next few problems are expectation and linearity of expectation problems. When finding the expected value of a random variable, first think about if the range is small enough so we can come up with the PMF and use the definition of expectation. Also, think about if there is a random variable from the zoo this random variable follows. If neither is possible, we will most likely want to use linearity Here's a general template for that!

1. **Decompose**. Write the random variable X as a sum of random variables: $X = X_1 + X_2 + ... + X_n$. Often, these X_i's are indicator random variables, especially if we're dealing with some kind of count.
2. **Apply LoE**. Apply LoE to $\mathbb{E}[X]$: $\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + ... + \mathbb{E}[X_n] = \sum_{i=1}^{n} \mathbb{E}[X_i]$.
3. **Conquer.** Compute each of $\mathbb{E}[X_i]$ and the plug it in to get the final answer.

Task 3 – Hungry Washing Machine

You have 10 pairs of socks (so 20 socks in total), with each pair being a different color. You put them in the washing machine, but the washing machine eats 4 of the socks chosen at random. Every subset of 4 socks is equally probable to be the subset that gets eaten. Let X be the number of complete pairs of socks that you have left.

a) What is the range of X, Ω_X (the set of possible values it can take on)? What is the probability mass function of X?

The washing machine eats 4 socks every time. It can either eat a single sock from 4 pairs of socks, leaving us with 6 complete pairs, or a single sock from 2 pairs and a matching pair, leaving us with 7 complete pairs, or 2 pairs of matching socks, leaving us with 8 complete pairs. That is,

$$\Omega_X = \{6, 7, 8\} .$$

We are dealing with a sample space with equally likely outcomes. As such, we can use the formula $\mathbb{P}(E) = \frac{|E|}{|\Omega|}$. We know that $|\Omega| = \binom{20}{4}$ because the washing machine picks a set of 4 socks out of 20 possible socks. To define the pmf of X, we consider each value in the range of X.

- For $k = 6$, we first pick 4 out of 10 pairs of socks from which we will eat a single sock ($\binom{10}{4}$ ways), and for each of these 4 pairs we have two socks to pick from ($\binom{2}{1}^4$ ways). Using the product rule, we get $|X = 6| = \binom{10}{4}2^4$.
- For $k = 7$, we first pick 1 out of 10 pairs of socks to eat in its entirety ($\binom{10}{1}$ ways), and then 2 out of the 9 remaining pairs from which we will eat a single sock ($\binom{9}{2}$ ways), and for each of these 2 pairs we have two socks to pick from ($\binom{2}{1}^2$ ways). Using the product rule, we get $|X = 7| = 10\binom{9}{2}2^2$.
- For $k = 8$, we pick 2 out of 10 pairs of socks to eat ($\binom{10}{2}$ ways). We get $|X = 8| = \binom{10}{2}$.

Thus,

$$p_X(k) = \begin{cases} \frac{\binom{10}{4}2^4}{\binom{20}{4}} & k = 6 \\ \frac{10\binom{9}{2}2^2}{\binom{20}{4}} & k = 7 \\ \frac{\binom{10}{2}}{\binom{20}{4}} & k = 8 \\ 0 & \text{otherwise} \end{cases}$$

b) Find $F_X(k)$, the CDF for X.

We can find the CDF by summing up the values in the PMF. For $k < 6$, the probability is 0 since there are no outcomes for these values. For $6k < 7$, the CDF is the probability of X being 6. For $7k < 8$, we add the probability of X being 7 to the previous cumulative probability. Finally, for $k8$, the CDF is 1 since we have included all possible outcomes.

$$F_X(k) = \begin{cases} 0 & \text{k¡6} \\ \frac{\binom{10}{4}2^4}{\binom{20}{4}} & 6 \leqslant k < 7 \\ \frac{\binom{10}{4}2^4}{\binom{20}{4}} + \frac{10\binom{9}{2}2^2}{\binom{20}{4}} & 7 \leqslant k < 8 \\ 1 & k \geqslant 8 \end{cases}$$

c) Find $\mathbb{E}[X]$ from the definition of expectation.

We calculate directly from the formula for expectation:

$$\mathbb{E}[X] = \sum_{k \in \Omega_X} k \cdot p_X(k) = 6 \cdot \frac{\binom{10}{4}2^4}{\binom{20}{4}} + 7 \cdot \frac{10\binom{9}{2}2^2}{\binom{20}{4}} + 8 \cdot \frac{\binom{10}{2}}{\binom{20}{4}} = \boxed{\frac{120}{19}}.$$

d) Find $\mathbb{E}[X]$ using linearity of expectation.

For $i \in [10]$, let X_i be 1 if pair i survived, and 0 otherwise. Then, $X = \sum_{i=1}^{10} X_i$. But $\mathbb{E}[X_i] = 1 \cdot \mathbb{P}(X_i = 1) + 0 \cdot \mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1) = \frac{\binom{18}{4}}{\binom{20}{4}}$, where the numerator indicates the number of ways of choosing 4 out the 18 remaining socks (we spare our chosen pair i). Hence,

$$\mathbb{E}[X] = \mathbb{E}[\sum_{i=1}^{10} X_i] = \sum_{i=1}^{10} \mathbb{E}[X_i] = \sum_{i=1}^{10} \frac{\binom{18}{4}}{\binom{20}{4}} = 10\frac{\binom{18}{4}}{\binom{20}{4}} = \boxed{\frac{120}{19}}$$

e) Which way was easier? Doing both (a) and (b), or just (c)?

Part (c) is was probably much easier. In this problem, you may have found part (a) and (b) easier, because there were only 3 possible values in the range of X. However, in general computing the probability mass function of complicated random variables (ones with hundreds of elements in their range) can be very difficult. Often it is much easier to use linearity of expectation and compute the probability mass function of simpler random variables.

Task 4 – 3-sided Die

Let the random variable X be the sum of two independent rolls of a fair 3-sided die. (If you are having trouble imagining what that looks like, you can use a 6-sided die and change the numbers on 3 of its faces.)

a) What is the probability mass function of X?

First let us define the range of X. A three sided-die can take on values $1, 2, 3$. Since X is the sum of two rolls, the range of X is $\Omega_X = \{2, 3, 4, 5, 6\}$.

We can then define the pmf of X. To that end, we must define two random variables R_1, R_2 with R_1 being the roll of the first die, and R_2 being the roll of the second die. Then, $X = R_1 + R_2$. Note that $\Omega_{R1} = \Omega_{R2} = \{1, 2, 3\}$. With that in mind we can find the pmf of X:

$$\begin{aligned} p_X(k) = \mathbb{P}(X = k) &= \sum_{i \in \Omega_{R1}} \mathbb{P}(R_1 = i, R_2 = k - i) \\ &= \sum_{i \in \Omega_{R1}} \mathbb{P}(R_1 = i) \cdot \mathbb{P}(R_2 = k - i) \qquad \text{(By independence of the rolls)} \\ &= \sum_{i \in \Omega_{R1}} \frac{1}{3} \cdot p_{R2}(k - i) \\ &= \frac{1}{3} \left(p_{R2}(k - 1) + p_{R2}(k - 2) + p_{R2}(k - 3) \right) \end{aligned}$$

At this point, we can evaluate the pmf of X for each value in the range of X, noting that $p_{R2}(k - i) = 0$ if $k - i \notin \Omega_{R2}$, $1/3$ otherwise. We get:

$$p_X(k) = \begin{cases} 1/9 & k = 2 \\ 2/9 & k = 3 \\ 3/9 & k = 4 \\ 2/9 & k = 5 \\ 1/9 & k = 6 \\ 0 & \text{otherwise} \end{cases}$$

One could also list out the possible values of the first two rolls and use a table to find the marginal pmf of X by summing up the entries of each row for each $k \in \Omega_X$.

b) What is the cumulative distribution function of X, partitioning the intervals on each possible value of X in its range?

Note that from part a), we know that the range of X is $\Omega_X = \{2, 3, 4, 5, 6\}$.
Remember that the CDF $F_X(k)$ is the probability that $X \leqslant k$. From the definition of CDF, the values that X can take, and the pmf of X, we get:

$$F_X(k) = \begin{cases} 0 & k < 2 \\ 1/9 & 2 \leqslant k < 3 \\ 3/9 & 3 \leqslant k < 4 \\ 6/9 & 4 \leqslant k < 5 \\ 8/9 & 5 \leqslant k < 6 \\ 1 & 6 \leqslant k \end{cases}$$

c) Find $\mathbb{E}[X]$ directly from the definition of expectation.

$$\mathbb{E}[X] = \sum_{k=2}^{6} k p_X(k) = 2 \cdot \frac{1}{9} + 3 \cdot \frac{2}{9} + 4 \cdot \frac{3}{9} + 5 \cdot \frac{2}{9} + 6 \cdot \frac{1}{9} = \boxed{4}$$

d) Find $\mathbb{E}[X]$ again, but this time using linearity of expectation.

Let R_1 be the roll of the first die, and R_2 the roll of the second. Then, $X = R_1 + R_2$.
By linearity of expectation, we get:

$$\mathbb{E}[X] = \mathbb{E}[R_1 + R_2] = \mathbb{E}[R_1] + \mathbb{E}[R_2]$$

We compute:

$$\mathbb{E}[R_1] = \sum_{i \in \Omega_{R1}} i \cdot \mathbb{P}(R_1 = i) = \sum_{i \in \Omega_{R1}} i \cdot \frac{1}{3} = \frac{1}{3}(1+2+3) = 2$$

Similarly, $\mathbb{E}[R_2] = 2$, since the rolls are independent.

Plugging into our expression for the expectation of X gives us:

$$\mathbb{E}[X] = 2 + 2 = \boxed{4}$$

Task 5 – Practice

a) Let X be a random variable with $p_X(k) = ck$ for $k \in \{1, \ldots, 5\} = \Omega_X$, and 0 otherwise. Find the value of c that makes X follow a valid probability distribution and compute its mean and variance ($\mathbb{E}[X]$ and $\mathrm{Var}(X)$).

For X to follow a valid probability distribution, we must have $\sum_{k \in \Omega_X} p_X(k) = 1$. We can solve for c so that the equality holds. We know:

$$\sum_{k \in \Omega_X} p_X(k) = \sum_{k \in \Omega_X} ck = c \sum_{k \in \Omega_X} k = c \cdot (1+2+3+4+5) = 15c$$

So for the normalization of the pmf of X to hold, we must choose $c = 1/15$.
We can now use the definition of expectation:

$$\mathbb{E}[X] = 1 \cdot \frac{1}{15} + 2 \cdot \frac{2}{15} + 3 \cdot \frac{3}{15} + 4 \cdot \frac{4}{15} + 5 \cdot \frac{5}{15} = 55/15 \approx \boxed{3.667}$$

And compute $\mathbb{E}[X]$ as follows:

$$\mathbb{E}[X^2] = 1^2 \cdot \frac{1}{15} + 2^2 \cdot \frac{2}{15} + 3^2 \cdot \frac{3}{15} + 4^2 \cdot \frac{4}{15} + 5^2 \cdot \frac{5}{15} = 225/15 = \boxed{15}$$

And the variance of X:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}^2[X] = 15 - (55/15)^2 = \frac{15^3 - 55^2}{15} = \frac{350}{225} = \frac{14}{9} \approx \boxed{1.556}$$

b) Let X be *any* random variable with mean $\mathbb{E}[X] = \mu$ and variance $\mathrm{Var}(X) = \sigma^2$. Find the mean and variance of $Z = \dfrac{X - \mu}{\sigma}$. (When you're done, you'll see why we call this a "standardized" version of X!)

We know that $\mathbb{E}[aX] = a \cdot \mathbb{E}[X]$ for some constant a, and that $\mathbb{E}[X + b] = \mathbb{E}[X] + b$ for some constant b. As such, we can compute the expectation of the standardized version of X, knowing that $\mathbb{E}[X] = \mu$:

$$\mathbb{E}[Z] = \mathbb{E}\left[\frac{X - \mu}{\sigma}\right] = \frac{1}{\sigma}\left(\mathbb{E}[X - \mu]\right) = \frac{1}{\sigma}(\mathbb{E}[X] - \mu) = \boxed{0}$$

For the variance, we know that $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$. With that in mind, knowing that $\mathrm{Var}(X) = \sigma^2$, we can write:

$$\mathrm{Var}(Z) = \mathrm{Var}\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma^2}\mathrm{Var}(X) = \boxed{1}$$

c) Let X, Y be independent random variables. Find the mean and variance of $X - 3Y - 5$ in terms of $\mathbb{E}[X], \mathbb{E}[Y]$, Var (X), and Var (Y).

Using the linearity of expectation, we can write:

$$\mathbb{E}[X - 3Y - 5] = \mathbb{E}[X] - 3\mathbb{E}[Y] - 5$$

We also know that the variance of a sum of independent random variables A and B is the sum of their variances, so that Var $(A + B)$ = Var (A) + Var (B). In our case, we have $A = X$, and $B = -3Y$. We get:

$$\text{Var}\,(X - 3Y - 5) = \text{Var}\,(X) + \text{Var}\,(-3Y) = \text{Var}\,(X) + 9\text{Var}\,(Y)$$

d) Let $X_1, \ldots, X_n$ be independent and identically distributed (iid) random variables each with mean μ and variance σ^2. The sample mean is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Find the mean and variance of $\bar{X}$. If you use the independence assumption anywhere, **explicitly label** at which step(s) it is necessary for your equalities to be true.

Using linearity of expectation,

$$\mathbb{E}[\overline{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}n\mu = \mu\,.$$

Note that independence was **not** necessary to calculate the above. As for variance,

$$\text{Var}\left(\overline{X}\right) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\,(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

where independence of the X_is is necessary for the second equality.

Task 6 – Symmetric Difference

For two sets A and B, define the **symmetric difference** Δ to be the set

$$A\Delta B = (A \cap B^C) \cup (B \cap A^C) = (A \cup B) \cap (A^C \cup B^C)\,,$$

i.e., the set containing elements that are in exactly one of A and B. For example, if $A = \{1,2,3\}$ and $B = \{2,3,4\}$, then $A\Delta B = \{1,4\}$, since 1 is in A and not in B, and 4 is in B and not in A. $2,3$ are in A and B, so they are not included in the symmetric difference.

Suppose A and B are random, independent (possibly empty) subsets of $\{1,2,\ldots,n\}$, where each subset is equally likely to be chosen as A or B. Let X be the random variable that is the size of $A\Delta B$ (in the example above, X would be 2). What is $\mathbb{E}[X]$?

For $i = 1,2,\ldots,n$, let X_i be the indicator of whether $i \in A\Delta B$. We may then say that $X = \sum_{i=1}^{n} X_i$. Note that if $i \in A\Delta B$, then i is either in A or i is in B. To that end, let Y_i and Z_i be the indicator variables of whether $i \in A$ and $i \in B$, respectively. Then

$$\mathbb{P}(Y_i = 1) = \frac{2^{n-1}}{2^n} = \frac{1}{2}\,,$$

where the numerator fixes i to be in A then constructs a subset from the remaining $n-1$ numbers, and the denominator is the number of subsets of $[n]$. A similar argument can be shown for $\mathbb{P}(Z_i = 1)$. Thus

$$\begin{aligned}\mathbb{E}[X_i] = \mathbb{P}(X_i = 1) &= \mathbb{P}(Y_i = 0, Z_i = 1) + \mathbb{P}(Y_i = 1, Z_i = 0)\\ &= \mathbb{P}(Y_i = 0)\,\mathbb{P}(Z_i = 1) + \mathbb{P}(Y_i = 1)\,\mathbb{P}(Z_i = 0)\\ &= \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{2}\,,\end{aligned}$$

where we have used the fact that Y_i, Z_i are independent. By Linearity of Expectation,

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n\cdot\frac{1}{2} = \frac{n}{2}\,.$$

Task 7 – Hat Check

At a reception, n people give their hats to a hat-check person. When they leave, the hat-check person gives each of them a hat chosen at random from the hats that remain. What is the expected number of people who get their own hats back? (Notice that the hats returned to two people are not independent events: if a certain hat is returned to one person, it cannot also be returned to the other person.)

Let X be the number of people who get their hats back. For $i \in [n]$, let X_i be 1 if person i gets their hat back, and 0 otherwise. Then, $\mathbb{E}[X_i] = \mathbb{P}(X_i = 1) = \frac{|E|}{|\Omega|}$. The sample space is all possible distributions of hats among the n people, and the event of interest E is the subset of the sample space where person i has their own hat. There are $n!$ ways to distribute the n hats among the n people. This is because the first person might have gotten 1 out of n possible hats; for each hat the first person got, the second person could get $n-1$ possible hats; and so on. The number of ways person i can get their hat back is $(n-1)!$. This is because we are essentially removing person i and hat i from the pool of people/hats, and counting the permutations of the $n-1$ remaining people.

Thus, $\mathbb{P}(X_i = 1) = \frac{(n-1)!}{n!} = \frac{1}{n}$. Since $X = \sum_{i=1}^{n} X_i$, Linearity of Expectation tell us that

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \frac{1}{n} = n\cdot\frac{1}{n} = 1\,.$$

Task 8 – Balls in Bins

Let X be the number of bins that remain empty when m balls are distributed into n bins randomly and independently. For each ball, each bin has an equal probability of being chosen. (Notice that two bins being empty are not independent events: if one bin is empty, that decreases the probability that the second bin will also be empty. This is particularly obvious when $n = 2$ and $m > 0$.) Find $\mathbb{E}[X]$.

For $i \in [n]$, let X_i be 1 if bin i is empty, and 0 otherwise. Then, $X = \sum_{i=1}^{n} X_i$. We first compute $\mathbb{E}[X_i] = 1 \cdot \mathbb{P}(X_i = 1) + 0 \cdot \mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1) = (\frac{n-1}{n})^m$. Indeed, we are assuming multiple balls can go in the same bin. As such, when computing $\mathbb{P}(X_i = 1)$, given that bin i is empty, we remove it from the pool of possible bins to pick from, leaving us with $n - 1$ bins out of a total of n bins in which we can place balls. Since we are distributing m balls over the n bins, the event that bin i remains empty occurs with probability $\left(\frac{n-1}{n}\right)^m$. Hence, by linearity of expectation:

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i\right] = n \cdot \left(\frac{n-1}{n}\right)^m$$

Task 9 – Frogger

A frog starts on a 1-dimensional number line at 0. At each second, independently, the frog takes a unit step right with probability p_1, to the left with probability p_2, and doesn't move with probability p_3, where $p_1 + p_2 + p_3 = 1$. After 2 seconds, let X be the location of the frog.

a) Find $p_X(k)$, the probability mass function for X.

Let L be a left step, R be a right step, and N be no step.
The range of X is $\{-2, -1, 0, 1, 2\}$. We can compute $p_X(-2) = \mathbb{P}(X = -2) = \mathbb{P}(LL) = p_2^2$, $p_X(-1) = \mathbb{P}(X = -1) = \mathbb{P}(LN \cup NL) = 2p_2p_3$, and $p_X(0) = \mathbb{P}(X = 0) = \mathbb{P}(NN \cup LR \cup RL) = p_3^2 + 2p_1p_2$. Similarly for $p_X(1)$ and $p_X(2)$.

$$p_X(k) = \begin{cases} p_2^2 & k = -2 \\ 2p_2p_3 & k = -1 \\ p_3^2 + 2p_1p_2 & k = 0 \\ 2p_1p_3 & k = 1 \\ p_1^2 & k = 2 \\ 0 & \text{otherwise} \end{cases}$$

b) Compute $\mathbb{E}[X]$ from the definition.

$$\mathbb{E}[X] = (-2)(p_2^2) + (-1)(2p_2p_3) + (0)(p_3^2 + 2p_1p_2) + (1)(2p_1p_3) + (2)(p_1^2) = 2(p_1 - p_2)$$

c) Compute $\mathbb{E}[X]$ again, but using linearity of expectation.

Let Y be the amount you moved on the first step (either $-1, 0, 1$), and Z the amount you moved on the second step. Then, $\mathbb{E}[Y] = \mathbb{E}[Z] = (1)(p_1) + (0)(p_3) + (-1)(p_2) = p_1 - p_2$.
Then $X = Y + Z$ and $\mathbb{E}[X] = \mathbb{E}[Y + Z] = \mathbb{E}[Y] + \mathbb{E}[Z] = 2(p_1 - p_2)$

Task 10 – Expectations, Independence, and Variance

a) Let U be a random variable which is uniform over the set $[n] = \{1, 2, \ldots, n\}$, i.e, $\mathbb{P}(U = i) = \frac{1}{n}$ for all $i \in [n]$. Compute $\mathbb{E}\left[U^2\right]$ and $\mathrm{Var}(U)$.

Hint: $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$ and $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$.

We first calculate $\mathbb{E}[U]$ directly from the definition of expectation:

$$\mathbb{E}[U] = \sum_{k=1}^{n} k \cdot p_U(k) = \sum_{k=1}^{n} k \cdot \frac{1}{n} = \frac{1}{n} \sum_{k=1}^{n} k = \frac{1}{n} \cdot \frac{n(n+1)}{2} .$$

We may calculate $\mathbb{E}[U^2]$, citing the Law of the Unconscious Statistician (LOTUS) with $g(X) = X^2$:

$$\mathbb{E}[U^2] = \sum_{k=1}^{n} g(k) \cdot p_U(k) = \sum_{k=1}^{n} k^2 \cdot p_U(k) = \frac{1}{n} \sum_{k=1}^{n} k^2 = \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6} .$$

Therefore

$$\begin{aligned} \mathrm{Var}(U) &= \mathbb{E}\left[U^2\right] - \mathbb{E}[U]^2 \\ &= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\ &= \frac{n+1}{12} \cdot (4n + 2 - 3n - 3) = \frac{(n+1)(n-1)}{12} . \end{aligned}$$

b) Let Y_1 and Y_2 be the independent outcomes of two fair 6-sided dice rolls, and let $Z = Y_1 + Y_2$. Then, compute $\mathbb{E}\left[Z^2\right]$ and $\mathrm{Var}(Z)$.

Hint: Try to use an indirect solution using linearity and independence, without the need of explicitly giving the distribution of Z^2.

First,

$$\begin{aligned} \mathbb{E}\left[Z^2\right] &= \mathbb{E}[(Y_1 + Y_2)^2] = \mathbb{E}[Y_1^2 + 2Y_1Y_2 + Y_2^2] \\ &= \mathbb{E}\left[Y_1^2\right] + \mathbb{E}\left[Y_2^2\right] + 2\mathbb{E}[Y_1 \cdot Y_2] && \text{linearity of expectation} \\ &= \mathbb{E}\left[Y_1^2\right] + \mathbb{E}\left[Y_2^2\right] + 2\mathbb{E}[Y_1]\,\mathbb{E}[Y_2] . && \text{independence} \end{aligned}$$

We know that $\mathbb{E}[Y_1] = \mathbb{E}[Y_2] = 21/6$. We also know that $\mathbb{E}\left[Y_1^2\right] = \mathbb{E}\left[Y_2^2\right] = 91/6$ (from class). Thus,

$$\mathbb{E}\left[Z^2\right] = 91/3 + 2 \cdot 21^2/36 = 91/3 + 147/6 = 329/6 .$$

On the other hand, we know that $\mathbb{E}[Z] = 7$. Therefore,

$$\mathrm{Var}(Z) = \mathbb{E}\left[Z^2\right] - \mathbb{E}[Z]^2 = 329/6 - 294/6 = 35/6 .$$

We could also have used $\mathrm{Var}(Z) = \mathrm{Var}(Y_1 + Y_2) = \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) = 35/12 \cdot 2 = 35/6$, using the calculation from class for the individual variances.

Task 11 – Pond fishing

Suppose I am fishing in a pond with B blue fish, R red fish, and G green fish, where $B + R + G = N$. For each of the following scenarios, identify the most appropriate distribution (with parameter(s)):

a) how many of the next 10 fish I catch are blue, if I catch and release

Since this is the same as saying how many of my next 10 trials (fish) are a success (are blue), this is a binomial distribution. Specifically, since we are doing catch and release, the probability of a given fish being blue is $\frac{B}{N}$ and each trial is independent. Thus:

$$\text{Bin}\left(10, \frac{B}{N}\right)$$

b) how many fish I had to catch until my first green fish, if I catch and release

Once again, each catch is independent, so this is asking how many trials until we see a success, hence it is a geometric distribution:

$$\text{Geo}\left(\frac{G}{N}\right)$$

c) how many red fish I catch in the next five minutes, if I catch on average r red fish per minute

This is asking for the number of occurrences of event given an average rate, which is the definition of the Poisson distribution. Since we're looking for events in the next 5 minutes, that is our time unit, so we have to adjust the average rate to match (r per minute becomes $5r$ per 5 minutes).

$$\text{Poi}(5r)$$

d) whether or not my next fish is blue

This is the same as the binomial case, but it's only one trial, so it is necessarily Bernoulli.

$$\text{Ber}\left(\frac{B}{N}\right)$$

e) how many of the next 10 fish I catch are blue, if I do not release the fish back to the pond after each catch

We have not covered the Hypergeometric RV in class, but its definition is the number of successes in n draws (without replacement) from N items that contain K successes in total. In this case, we have 10 draws (without replacement because we do not catch and release), and out of the N fish, B are blue (a success).

$$\text{HypGeo}(N, B, 10)$$

f) how many fish I have to catch until I catch three red fish, if I catch and release

Negative binomial is another RV we didn't cover in class. It models the number of trials with probability of success p, until you get r successes. In this case, as before, our trials are caught fish (with replacement this time) and our success is if the fish are red, which happens with probability $\frac{R}{N}$.

$$\text{NegBin}\left(3, \frac{R}{N}\right)$$