Chapter 3: Expectation and Variance

In the previous chapter we looked at probability, with three major themes:

1. Conditional probability: $\mathbb{P}(A \mid B)$.
2. First-step analysis for calculating eventual probabilities in a stochastic process.
3. Calculating probabilities for continuous and discrete random variables.

In this chapter, we look at the same themes for **expectation** and **variance**. The expectation of a random variable is the *long-term average of the random variable.*

Imagine observing many thousands of independent random values from the random variable of interest. Take the average of these random values. The expectation is the value of this average as the sample size tends to infinity.

We will repeat the three themes of the previous chapter, but in a different order.

1. Calculating expectations for continuous and discrete random variables.
2. Conditional expectation: the expectation of a random variable X, *conditional* on the value taken by another random variable Y. If the value of Y affects the value of X (i.e. X and Y are *dependent*), the conditional expectation of X given the value of Y will be different from the overall expectation of X.
3. First-step analysis for calculating the expected amount of time needed to reach a particular state in a process (e.g. the expected number of shots before we win a game of tennis).

We will also study similar themes for variance.

3.1 Expectation

The **mean**, **expected value**, or **expectation** of a random variable X is written as $\mathbb{E}(X)$ or μ_X. If we observe N random values of X, then the mean of the N values will be approximately equal to $\mathbb{E}(X)$ for large N. The expectation is defined differently for continuous and discrete random variables.

Definition: Let X be a **continuous** random variable with p.d.f. $f_X(x)$. The expected value of X is

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

Definition: Let X be a **discrete** random variable with probability function $f_X(x)$. The expected value of X is

$$\mathbb{E}(X) = \sum_x x f_X(x) \ = \sum_x x\mathbb{P}(X = x).$$

Expectation of $g(X)$

Let $g(X)$ be a function of X. We can imagine a long-term average of $g(X)$ just as we can imagine a long-term average of X. This average is written as $\mathbb{E}(g(X))$. Imagine observing X many times (N times) to give results $x_1, x_2, \ldots, x_N$. Apply the function g to each of these observations, to give $g(x_1), \ldots, g(x_N)$. The mean of $g(x_1), g(x_2), \ldots, g(x_N)$ approaches $\mathbb{E}(g(X))$ as the number of observations N tends to infinity.

Definition: Let X be a **continuous** random variable, and let g be a function. The expected value of $g(X)$ is

$$\mathbb{E}\Big(g(X)\Big) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx.$$

Definition: Let X be a **discrete** random variable, and let g be a function. The expected value of $g(X)$ is

$$\mathbb{E}\Big(g(X)\Big) = \sum_x g(x) f_X(x) = \sum_x g(x)\mathbb{P}(X = x).$$

Expectation of XY: the definition of $\mathbb{E}(XY)$

Suppose we have two random variables, X and Y. These might be independent, in which case the value of X has no effect on the value of Y. Alternatively, X and Y might be *dependent*: when we observe a random value for X, it might influence the random values of Y that we are most likely to observe. For example, X might be the height of a randomly selected person, and Y might be the weight. On the whole, larger values of X will be associated with larger values of Y.

To understand what $\mathbb{E}(XY)$ means, think of observing a large number of *pairs* $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$. If X and Y are dependent, the value x_i might affect the value y_i, and vice versa, so we have to keep the observations together in their pairings. As the number of pairs N tends to infinity, the average $\frac{1}{N}\sum_{i=1}^{N} x_i \times y_i$ approaches the expectation $\mathbb{E}(XY)$.

For example, if X is height and Y is weight, $\mathbb{E}(XY)$ is the average of (height $\times$ weight). We are interested in $\mathbb{E}(XY)$ because it is used for calculating the *covariance* and *correlation*, which are measures of how closely related X and Y are (see Section 3.2).

Properties of Expectation

i) Let g and h be functions, and let a and b be constants. For any random variable X (discrete or continuous),

$$\mathbb{E}\Big\{ag(X) + bh(X)\Big\} = a\mathbb{E}\Big\{g(X)\Big\} + b\mathbb{E}\Big\{h(X)\Big\}.$$

In particular,

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b.$$

ii) Let X and Y be ANY random variables (discrete, continuous, independent, or non-independent). Then

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

More generally, for ANY random variables $X_1, \ldots, X_n$,

$$\mathbb{E}(X_1 + \ldots + X_n) = \mathbb{E}(X_1) + \ldots + \mathbb{E}(X_n).$$

iii) Let X and Y be **independent** random variables, and g, h be functions. Then

$$\begin{aligned} \mathbb{E}(XY) &= \mathbb{E}(X)\mathbb{E}(Y) \\ \mathbb{E}\Big(g(X)h(Y)\Big) &= \mathbb{E}\Big(g(X)\Big)\mathbb{E}\Big(h(Y)\Big). \end{aligned}$$

Notes: 1. $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ is ONLY generally true if X and Y are *INDEPENDENT.*

2. If X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. However, the converse is not generally true: it is possible for $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ even though X and Y are dependent.

Probability as an Expectation

Let A be any event. We can write $\mathbb{P}(A)$ as an expectation, as follows.
Define the ***indicator function:***

$$I_A = \begin{cases} 1 & \textit{if event } A \textit{ occurs,} \\ 0 & \textit{otherwise.} \end{cases}$$

Then I_A is a ***random variable***, and

$$\begin{aligned} \mathbb{E}(I_A) &= \sum_{r=0}^{1} r\mathbb{P}(I_A = r) \\ &= 0 \times \mathbb{P}(I_A = 0) + 1 \times \mathbb{P}(I_A = 1) \\ &= \mathbb{P}(I_A = 1) \\ &= \mathbb{P}(A). \end{aligned}$$

Thus

$$\boxed{\mathbb{P}(A) = \mathbb{E}(I_A) \textit{ for any event } A.}$$

3.2 Variance, covariance, and correlation

The variance of a random variable X is a measure of how ***spread out*** it is. Are the values of X clustered tightly around their mean, or can we commonly observe values of X a long way from the mean value? The *variance* measures how far the values of X are from their mean, on average.

Definition: Let X be any random variable. The **variance** of X is

$$\mathsf{Var}(X) = \mathbb{E}\Big((X-\mu_X)^2\Big) = \mathbb{E}(X^2) - \Big(\mathbb{E}(X)\Big)^2.$$

The variance is the ***mean squared deviation*** of a random variable from its own mean.

If X has *high variance*, we can observe values of X a long way from the mean.

If X has *low variance*, the values of X tend to be clustered tightly around the mean value.

Example: Let X be a continuous random variable with p.d.f.

$$f_X(x) = \begin{cases} 2x^{-2} & \text{for } 1 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Find $\mathbb{E}(X)$ and $\mathrm{Var}(X)$.

$$\begin{aligned} E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx = \int_1^2 x \times 2x^{-2}\, dx &= \int_1^2 2x^{-1}\, dx \\ &= \Big[\, 2\log(x) \Big]_1^2 \\ &= 2\log(2) - 2\log(1) \\ &= 2\log(2). \end{aligned}$$

For $\mathrm{Var}(X)$*, we use*

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2 \ .$$

Now

$$\begin{aligned} E(X^2) = \int_{-\infty}^{\infty} x^2 \, f_X(x)\, dx = \int_1^2 x^2 \times 2x^{-2}\, dx &= \int_1^2 2\, dx \\ &= \Big[2x\Big]_1^2 \\ &= 2 \times 2 - 2 \times 1 \\ &= 2. \end{aligned}$$

Thus

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2 \\ &= 2 - \{2\log(2)\}^2 \\ &= 0.0782. \end{aligned}$$

Covariance

Covariance is a measure of the association or dependence between two random variables X and Y. Covariance can be either positive or negative. (*Variance* is always positive.)

Definition: Let X and Y be any random variables. The **covariance** between X and Y is given by

$$\mathrm{cov}(X,Y) = \mathbb{E}\Big\{(X-\mu_X)(Y-\mu_Y)\Big\} = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y),$$

where $\mu_X = \mathbb{E}(X),\ \mu_Y = \mathbb{E}(Y)$.

1. $\mathrm{cov}(X,Y)$ will be ***positive*** if large values of X tend to occur with large values of Y, and small values of X tend to occur with small values of Y. For example, if X is height and Y is weight of a randomly selected person, we would expect $\mathrm{cov}(X,Y)$ to be positive.

2. $\text{cov}(X, Y)$ will be ***negative*** if large values of X tend to occur with small values of Y, and small values of X tend to occur with large values of Y. For example, if X is age of a randomly selected person, and Y is heart rate, we would expect X and Y to be negatively correlated (older people have slower heart rates).

3. If X and Y are independent, then there is no pattern between large values of X and large values of Y, so $\text{cov}(X, Y) = 0$. However, $\text{cov}(X, Y) = 0$ does NOT imply that X and Y are independent, unless X and Y are Normally distributed.

Properties of Variance

i) Let g be a function, and let a and b be constants. For any random variable X (discrete or continuous),

$$\textit{Var}\Big\{ag(X) + b\Big\} = a^2\,\textit{Var}\Big\{g(X)\Big\}.$$

In particular, $\textit{Var}(aX + b) = a^2\,\textit{Var}(X)$.

ii) Let X and Y be **independent** random variables. Then

$$\textit{Var}(X + Y) = \textit{Var}(X) + \textit{Var}(Y).$$

iii) If X and Y are **NOT independent**, then

$$\textit{Var}(X + Y) = \textit{Var}(X) + \textit{Var}(Y) + 2\textit{cov}(X, Y).$$

Correlation (non-examinable)

The correlation coefficient of X and Y is a measure of the linear association between X and Y. It is given by the covariance, scaled by the overall variability in X and Y. As a result, the correlation coefficient is always between -1 and $+1$, so it is easily compared for different quantities.

Definition: The **correlation** between X and Y, also called the **correlation coefficient**, is given by

$$\boxed{\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.}$$

The correlation measures linear association between X and Y. It takes values only between -1 and $+1$, and has the same sign as the covariance.

The correlation is ± 1 if and only if there is a perfect linear relationship between X and Y, i.e. $\text{corr}(X, Y) = 1 \iff Y = aX + b$ for some constants a and b.

The correlation is 0 if X and Y are independent, but a correlation of 0 does not *imply* that X and Y are independent.

3.3 Conditional Expectation and Conditional Variance

Throughout this section, we will assume for simplicity that X and Y are discrete random variables. However, exactly the same results hold for continuous random variables too.

Suppose that X and Y are discrete random variables, possibly dependent on each other. Suppose that we fix Y at the value y. This gives us a set of conditional probabilities $\mathbb{P}(X = x \,|\, Y = y)$ for all possible values x of X. This is called the ***conditional distribution of X, given that $Y = y$.***

Definition: Let X and Y be discrete random variables. The **conditional probability function** of X, given that $Y = y$, is:

$$\mathbb{P}(X = x \,|\, Y = y) = \frac{\mathbb{P}(X = x \textit{ AND } Y = y)}{\mathbb{P}(Y = y)}.$$

We write the conditional probability function as:

$$f_{X\,|\,Y}(x \,|\, y) = \mathbb{P}(X = x \,|\, Y = y).$$

Note: The conditional probabilities $f_{X\,|\,Y}(x \,|\, y)$ sum to one, just like any other probability function:

$$\sum_x \mathbb{P}(X = x \,|\, Y = y) = \sum_x \mathbb{P}_{\{Y=y\}}(X = x) = 1,$$

using the subscript notation $\mathbb{P}_{\{Y=y\}}$ of Section 2.3.

We can also find the expectation and variance of X with respect to this conditional distribution. That is, if we know that the value of Y is fixed at y, then we can find the mean value of X *given that* Y takes the value y, and also the variance of X given that $Y = y$.

Definition: Let X and Y be discrete random variables. The **conditional expectation of X, given that $Y = y$**, is

$$\mu_{X\,|\,Y=y} = \mathbb{E}(X \,|\, Y = y) = \sum_x x f_{X\,|\,Y}(x \,|\, y).$$

$\mathbb{E}(X \,|\, Y = y)$ is the *mean value of X, when Y is fixed at y.*

Conditional expectation as a random variable

The unconditional expectation of X, $\mathbb{E}(X)$, is just *a number:*
e.g. $\mathbb{E}X = 2$ *or* $\mathbb{E}X = 5.8$.

The conditional expectation, $\mathbb{E}(X \,|\, Y = y)$, is *a number depending on y.*

If Y has an influence on the value of X, then Y will have an influence on the *average* value of X. So, for example, we would expect $\mathbb{E}(X \,|\, Y = 2)$ to be different from $\mathbb{E}(X \,|\, Y = 3)$.

We can therefore view $\mathbb{E}(X \,|\, Y = y)$ as a *function of y, say* $\mathbb{E}(X \,|\, Y{=}y) = h(y)$.

To evaluate this function, $h(y) = \mathbb{E}(X \,|\, Y = y)$, we:

i) *fix Y at the chosen value y;*

ii) *find the expectation of X when Y is fixed at this value.*

However, we could also evaluate the function at a *random value* of Y:

i) *observe a random value of* Y*;*

ii) *fix* Y *at that observed random value;*

iii) *evaluate* $\mathbb{E}(X \mid Y =$ *observed random value*$)$.

We obtain a *random variable*: $\mathbb{E}(X \mid Y) = h(Y)$.

The randomness comes from the randomness in Y*, not in* X.

> *Conditional expectation,* $\mathbb{E}(X \mid Y)$*, is a random variable with randomness inherited from* Y*, not* X.

Example: Suppose $Y = \begin{cases} 1 & \text{with probability } 1/8\,, \\ 2 & \text{with probability } 7/8\,, \end{cases}$

and $X \mid Y = \begin{cases} 2Y & \text{with probability } 3/4\,, \\ 3Y & \text{with probability } 1/4\,. \end{cases}$

Conditional expectation of X given $Y = y$ is a number depending on y:

If $Y = 1$*, then:* $X \mid (Y = 1) = \begin{cases} 2 & \textit{with probability } 3/4 \\ 3 & \textit{with probability } 1/4 \end{cases}$

$$\textit{so} \quad \mathbb{E}(X \mid Y = 1) = 2 \times \tfrac{3}{4} + 3 \times \tfrac{1}{4} = \tfrac{9}{4}.$$

If $Y = 2$*, then:* $X \mid (Y = 2) = \begin{cases} 4 & \textit{with probability } 3/4 \\ 6 & \textit{with probability } 1/4 \end{cases}$

$$\textit{so} \quad \mathbb{E}(X \mid Y = 2) = 4 \times \tfrac{3}{4} + 6 \times \tfrac{1}{4} = \tfrac{18}{4}.$$

Thus $\mathbb{E}(X \mid Y = y) = \begin{cases} 9/4 & \textit{if } y = 1 \\ 18/4 & \textit{if } y = 2. \end{cases}$

So $\mathbb{E}(X \mid Y = y)$ *is a number depending on* y*, or a function of* y.

Conditional expectation of X given random Y is a random variable:

From above, $\mathbb{E}(X \mid Y) = \begin{cases} 9/4 & \text{if } Y = 1 \text{ (probability } 1/8), \\ 18/4 & \text{if } Y = 2 \text{ (probability } 7/8). \end{cases}$

So $\mathbb{E}(X \mid Y) = \begin{cases} 9/4 & \text{with probability } 1/8, \\ 18/4 & \text{with probability } 7/8. \end{cases}$

Thus $\mathbb{E}(X \mid Y)$ *is a random variable.*

The randomness in $\mathbb{E}(X \mid Y)$ *is inherited from* Y*, not from* X*.*

Conditional expectation is a very useful tool for finding the ***unconditional*** expectation of X (see below). Just like the Partition Theorem, it is useful because it is often easier to specify conditional probabilities than to specify overall probabilities.

Conditional variance

The conditional variance is similar to the conditional expectation.

- $\mathrm{Var}(X \mid Y = y)$ is the variance of X, when Y is fixed at the value $Y = y$.
- $\mathrm{Var}(X \mid Y)$ is a random variable, giving the variance of X when Y is fixed at a value to be selected randomly.

Definition: Let X and Y be random variables. The **conditional variance of X, given Y,** is given by

$$\mathrm{Var}(X \mid Y) = \mathbb{E}(X^2 \mid Y) - \Big\{\mathbb{E}(X \mid Y)\Big\}^2 = \mathbb{E}\Big\{(X - \mu_{X \mid Y})^2 \mid Y\Big\}$$

Like expectation, $\mathrm{Var}(X \mid Y = y)$ *is a number depending on* y *(a function of* y*), while* $\mathrm{Var}(X \mid Y)$ *is a random variable with randomness inherited from* Y*.*

Laws of Total Expectation and Variance

If all the expectations below are finite, then for ANY random variables X and Y, we have:

i) $\boxed{\mathbb{E}(X) = \mathbb{E}_Y\Big(\mathbb{E}(X \mid Y)\Big)}$ *Law of Total Expectation.*

Note that we can pick any r.v. Y, to make the expectation as easy as we can.

ii) $\mathbb{E}(g(X)) = \mathbb{E}_Y\Big(\mathbb{E}(g(X) \mid Y)\Big)$ *for any function g.*

iii) $\boxed{\mathit{Var}(X) = \mathbb{E}_Y\Big(\mathit{Var}(X \mid Y)\Big) + \mathit{Var}_Y\Big(\mathbb{E}(X \mid Y)\Big)}$

Law of Total Variance.

Note: $\mathbb{E}_Y$ and Var_Y denote expectation over Y and variance over Y,

i.e. the expectation or variance is computed over the distribution of the random variable Y.

The Law of Total Expectation says that *the total average is the average of case-by-case averages.*

- The total average is $\mathbb{E}(X)$*;*
- The case-by-case averages are $\mathbb{E}(X \mid Y)$ *for the different values of Y;*
- The average of case-by-case averages is *the average over Y of the Y-case averages:* $\mathbb{E}_Y\Big(\mathbb{E}(X \mid Y)\Big)$.

Example: In the example above, we had: $\mathbb{E}(X \,|\, Y) = \begin{cases} 9/4 & \text{with probability } 1/8, \\ 18/4 & \text{with probability } 7/8. \end{cases}$

The total average is:

$$\mathbb{E}(X) = \mathbb{E}_Y\Big\{\mathbb{E}(X \,|\, Y)\Big\} = \frac{9}{4} \times \frac{1}{8} + \frac{18}{4} \times \frac{7}{8} = 4.22.$$

Proof of (i), (ii), (iii):

(i) is a special case of (ii), so we just need to prove (ii). Begin at RHS:

$$\begin{aligned}
\text{RHS} \;=\; \mathbb{E}_Y\Big[\mathbb{E}(g(X) \,|\, Y)\Big] \;&=\; \mathbb{E}_Y\left[\sum_x g(x)\mathbb{P}(X = x \,|\, Y)\right] \\
&=\; \sum_y \left[\sum_x g(x)\mathbb{P}(X = x \,|\, Y = y)\right]\mathbb{P}(Y = y) \\
&=\; \sum_y \sum_x g(x)\mathbb{P}(X = x \,|\, Y = y)\mathbb{P}(Y = y) \\
&=\; \sum_x g(x) \sum_y \mathbb{P}(X = x \,|\, Y = y)\mathbb{P}(Y = y) \\
&=\; \sum_x g(x)\mathbb{P}(X = x) \qquad \text{(partition rule)} \\
&=\; \mathbb{E}(g(X)) \quad = \quad \text{LHS.}
\end{aligned}$$

(iii) Wish to prove $\text{Var}(X) = \mathbb{E}_Y[\text{Var}(X \,|\, Y)] + \text{Var}_Y[\mathbb{E}(X \,|\, Y)]$. Begin at RHS:

$$\mathbb{E}_Y[\text{Var}(X \,|\, Y)] + \text{Var}_Y[\mathbb{E}(X \,|\, Y)]$$

$$= \mathbb{E}_Y\Big\{\mathbb{E}(X^2 \,|\, Y) - (\mathbb{E}(X \,|\, Y))^2\Big\} + \left\{\mathbb{E}_Y\Big\{[\mathbb{E}(X \,|\, Y)]^2\Big\} - \Big[\underbrace{\mathbb{E}_Y(\mathbb{E}(X \,|\, Y))}_{\mathbb{E}(X) \text{ by part (i)}}\Big]^2\right\}$$

$$= \underbrace{\mathbb{E}_Y\{\mathbb{E}(X^2 \,|\, Y)\}}_{\mathbb{E}(X^2) \text{ by part (i)}} - \mathbb{E}_Y\{[\mathbb{E}(X \,|\, Y)]^2\} + \mathbb{E}_Y\{[\mathbb{E}(X \,|\, Y)]^2\} - (\mathbb{E}X)^2$$

$$= \mathbb{E}(X^2) - (\mathbb{E}X)^2$$

$$= \text{Var}(X) = \text{LHS.} \qquad \square$$

3.4 Examples of Conditional Expectation and Variance

1. Swimming with dolphins

Fraser runs a dolphin-watch business. Every day, he is unable to run the trip due to bad weather with probability p, independently of all other days. Fraser works every day except the bad-weather days, which he takes as holiday.

Let Y be the number of consecutive days Fraser has to work between bad-weather days. Let X be the total number of customers who go on Fraser's trip in this period of Y days. Conditional on Y, the distribution of X is

$$(X \mid Y) \sim \text{Poisson}(\mu Y).$$

(a) Name the distribution of Y, and state $\mathbb{E}(Y)$ and $\text{Var}(Y)$.

(b) Find the expectation and the variance of the number of customers Fraser sees between bad-weather days, $\mathbb{E}(X)$ and $\text{Var}(X)$.

(a) *Let 'success' be 'bad-weather day' and 'failure' be 'work-day'.*
Then $\mathbb{P}(\textit{success}) = \mathbb{P}(\textit{bad-weather}) = p$.
Y *is the number of failures before the first success.*
So

$$Y \sim \textit{Geometric}(p).$$

Thus

$$\mathbb{E}(Y) = \frac{1-p}{p},$$

$$\textit{Var}(Y) = \frac{1-p}{p^2}.$$

(b) *We know* $(X \mid Y) \sim \textit{Poisson}(\mu Y)$*: so*

$$\mathbb{E}(X \mid Y) = \textit{Var}(X \mid Y) = \mu Y.$$

By the Law of Total Expectation:

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}_Y\Big\{\mathbb{E}(X \mid Y)\Big\} \\ &= \mathbb{E}_Y(\mu Y) \\ &= \mu\mathbb{E}_Y(Y)\end{aligned}$$

$$\therefore \quad \mathbb{E}(X) = \frac{\mu(1-p)}{p}\,.$$

By the Law of Total Variance:

$$\begin{aligned}\mathit{Var}(X) &= \mathbb{E}_Y\Big(\mathit{Var}(X \mid Y)\Big) + \mathit{Var}_Y\Big(\mathbb{E}(X \mid Y)\Big) \\ &= \mathbb{E}_Y\Big(\mu Y\Big) + \mathit{Var}_Y\Big(\mu Y\Big) \\ &= \mu\mathbb{E}_Y(Y) + \mu^2\,\mathit{Var}_Y(Y) \\ &= \mu\left(\frac{1-p}{p}\right) + \mu^2\left(\frac{1-p}{p^2}\right) \\ &= \frac{\mu(1-p)(p+\mu)}{p^2}\,.\end{aligned}$$

Checking your answer in R:

If you know how to use a statistical package like R, you can check your answer to the question above as follows.

```
> # Pick a value for p, e.g. p = 0.2.
> # Pick a value for mu, e.g. mu = 25
>
> # Generate 10,000 random values of Y ~ Geometric(p = 0.2):
> y <- rgeom(10000, prob=0.2)
>
> # Generate 10,000 random values of X conditional on Y:
> # use (X | Y) ~ Poisson(mu * Y) ~ Poisson(25 * Y)
> x <- rpois(10000, lambda = 25*y)
```

```
> # Find the sample mean of X (should be close to E(X)):
> mean(x)
[1] 100.6606
>
> # Find the sample variance of X (should be close to var(X)):
> var(x)
[1] 12624.47
>
> # Check the formula for E(X):
> 25 * (1 - 0.2) / 0.2
[1] 100
>
> # Check the formula for var(X):
> 25 * (1 - 0.2) * (0.2 + 25) / 0.2^2
[1] 12600
```

The formulas we obtained by working give $\mathbb{E}(X) = 100$ and $\mathrm{Var}(X) = 12600$. The sample mean was $\overline{x} = 100.6606$ (close to 100), and the sample variance was 12624.47 (close to 12600). Thus our working seems to have been correct.

---

2. Randomly stopped sum

This model arises very commonly in stochastic processes. A random number N of events occur, and each event i has associated with it some cost, penalty, or reward X_i. The question is to find the mean and variance of the total cost / reward:
$T_N = X_1 + X_2 + \ldots + X_N$.
The difficulty is that the number N of terms in the sum is itself random.

T_N is called a *randomly stopped sum: it is a sum of X_i's, randomly stopped at the random number of N terms.*

Example: Think of a cash machine, which has to be loaded with enough money to cover the day's business. The number of customers per day is a random number N. Customer i withdraws a random amount X_i. The total amount withdrawn during the day is a randomly stopped sum: $T_N = X_1 + \ldots + X_N$.

Cash machine example

The citizens of Remuera withdraw money from a cash machine according to the following probability function (X):

Amount, x (\$)	50	100	200
$\mathbb{P}(X = x)$	0.3	0.5	0.2

The number of customers per day has the distribution $N \sim \text{Poisson}(\lambda)$.

Let $T_N = X_1 + X_2 + \ldots + X_N$ be the total amount of money withdrawn in a day, where each X_i has the probability function above, and $X_1, X_2, \ldots$ are independent of each other and of N.
T_N is a randomly stopped sum, stopped by the random number of N customers.

(a) Show that $\mathbb{E}(X) = 105$, and $\text{Var}(X) = 2725$.

(b) Find $\mathbb{E}(T_N)$ and $\text{Var}(T_N)$: the mean and variance of the amount of money withdrawn each day.

Solution

(a) Exercise.

(b) *Let $T_N = \sum_{i=1}^{N} X_i$. If we knew how many terms were in the sum, we could easily find $\mathbb{E}(T_N)$ and Var(T_N) as the mean and variance of a sum of independent r.v.s.*
So 'pretend' we know how many terms are in the sum: i.e. condition on N.

$$\begin{aligned}
\mathbb{E}(T_N \mid N) &= \mathbb{E}(X_1 + X_2 + \ldots + X_N \mid N) \\
&= \mathbb{E}(X_1 + X_2 + \ldots + X_N) \\
&\qquad \textit{(because all } X_i\textit{s are independent of } N\textit{)} \\
&= \mathbb{E}(X_1) + \mathbb{E}(X_2) + \ldots + \mathbb{E}(X_N) \\
&\qquad \textit{where } N \textit{ is now considered constant;} \\
&\qquad \textit{(we do NOT need independence of } X_i\textit{'s for this)} \\
&= N \times \mathbb{E}(X) \quad \textit{(because all } X_i\textit{'s have same mean, } \mathbb{E}(X)\textit{)} \\
&= 105N.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\textit{Var}(T_N \mid N) &= \textit{Var}(X_1 + X_2 + \ldots + X_N \mid N) \\
&= \textit{Var}(X_1 + X_2 + \ldots + X_N) \\
&\qquad \textit{where } N \textit{ is now considered constant;} \\
&\qquad \textit{(because all } X_i\textit{'s are independent of } N\textit{)} \\
&= \textit{Var}(X_1) + \textit{Var}(X_2) + \ldots + \textit{Var}(X_N) \\
&\qquad \textit{(we DO need independence of } X_i\textit{'s for this)} \\
&= N \times \textit{Var}(X) \quad \textit{(because all } X_i\textit{'s have same variance, } \textit{Var}(X)\textit{)} \\
&= 2725N.
\end{aligned}$$

So

$$\begin{aligned}
\mathbb{E}(T_N) &= \mathbb{E}_N \Big\{ \mathbb{E}(T_N \mid N) \Big\} \\
&= \mathbb{E}_N(105N) \\
&= 105\mathbb{E}_N(N) \\
&= 105\lambda,
\end{aligned}$$

because $N \sim$ *Poisson*(λ) *so* $\mathbb{E}(N) = \lambda$*.*

Similarly,

$$\begin{aligned}
\textit{Var}(T_N) &= \mathbb{E}_N \Big\{ \textit{Var}(T_N \mid N) \Big\} + \textit{Var}_N \Big\{ \mathbb{E}(T_N \mid N) \Big\} \\
&= \mathbb{E}_N \{2725N\} + \textit{Var}_N \{105N\} \\
&= 2725\mathbb{E}_N(N) + 105^2 \textit{Var}_N(N) \\
&= 2725\lambda + 11025\lambda \\
&= 13750\lambda,
\end{aligned}$$

because $N \sim$ *Poisson*(λ) *so* $\mathbb{E}(N) = \textit{Var}(N) = \lambda$*.*

Check in R (advanced)

```
> # Create a function tn.func to calculate a single value of T_N
> # for a given value N=n:
> tn.func <- function(n){
         sum(sample(c(50, 100, 200), n, replace=T,
         prob=c(0.3, 0.5, 0.2)))
   }

> # Generate 10,000 random values of N, using lambda=50:
> N <- rpois(10000, lambda=50)
> # Generate 10,000 random values of T_N, conditional on N:
> TN <- sapply(N, tn.func)
> # Find the sample mean of T_N values, which should be close to
> # 105 * 50 = 5250:
> mean(TN)
[1] 5253.255
> # Find the sample variance of T_N values, which should be close
> # to 13750 * 50 = 687500:
> var(TN)
[1] 682469.4
```

All seems well. Note that the sample variance is often some distance from the true variance, even when the sample size is 10,000.

General result for randomly stopped sums:

Suppose $X_1, X_2, \ldots$ each have the same mean μ and variance σ^2, and $X_1, X_2, \ldots$, and N are mutually independent. Let $T_N = X_1 + \ldots + X_N$ be the randomly stopped sum. By following similar working to that above:

$$\mathbb{E}(T_N) = \mathbb{E}\left\{\sum_{i=1}^{N} X_i\right\} = \mu\,\mathbb{E}(N)$$

$$\mathrm{Var}(T_N) = \mathrm{Var}\left\{\sum_{i=1}^{N} X_i\right\} = \sigma^2\,\mathbb{E}(N) + \mu^2\,\mathrm{Var}(N).$$

3.5 First-Step Analysis for calculating expected reaching times

Remember from Section 2.6 that we use First-Step Analysis for finding the probability of eventually reaching a particular state in a stochastic process. First-step analysis for probabilities uses *conditional probability and the Partition Theorem (Law of Total Probability).*

In the same way, we can use first-step analysis for finding the *expected reaching time for a state.*

This is the expected number of steps that will be needed to reach a particular state from a specified start-point, or the expected length of time it will take to get there if we have a continuous time process.

Just as first-step analysis for probabilities uses conditional probability and the law of total probability (Partition Theorem), first-step analysis for expectations uses *conditional expectation and the law of total expectation.*

First-step analysis for probabilities:

The first-step analysis procedure for probabilities can be summarized as follows:

$$\mathbb{P}(\textit{eventual goal}) = \sum_{\substack{\textit{first-step} \\ \textit{options}}} \mathbb{P}(\textit{eventual goal} \,|\, \textit{option})\mathbb{P}(\textit{option}) \,.$$

This is because the first-step options form a *partition of the sample space.*

First-step analysis for expected reaching times:

The expression for expected reaching times is very similar:

$$\mathbb{E}(\textit{reaching time}) = \sum_{\substack{\textit{first-step} \\ \textit{options}}} \mathbb{E}(\textit{reaching time} \,|\, \textit{option})\mathbb{P}(\textit{option}) \,.$$

This follows immediately from the law of total expectation:

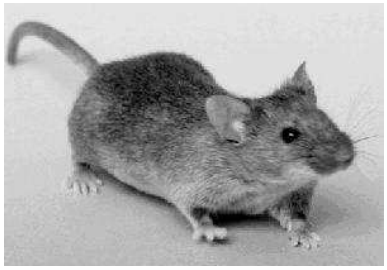
$$\mathbb{E}(X) = \mathbb{E}_Y\Big\{\mathbb{E}(X \mid Y)\Big\} = \sum_y \mathbb{E}(X \mid Y = y)\mathbb{P}(Y = y).$$

Let X be the reaching time, and let Y be the label for possible options:
i.e. $Y = 1, 2, 3, \ldots$ *for options 1, 2, 3, ...*

We then obtain:

$$\begin{aligned} \mathbb{E}(X) &= \sum_y \mathbb{E}(X \mid Y = y)\mathbb{P}(Y = y) \\ \textit{i.e.} \quad \mathbb{E}(\textit{reaching time}) &= \sum_{\substack{\textit{first-step} \\ \textit{options}}} \mathbb{E}(\textit{reaching time} \mid \textit{option})\mathbb{P}(\textit{option}) \,. \end{aligned}$$

Example 1: Mouse in a Maze



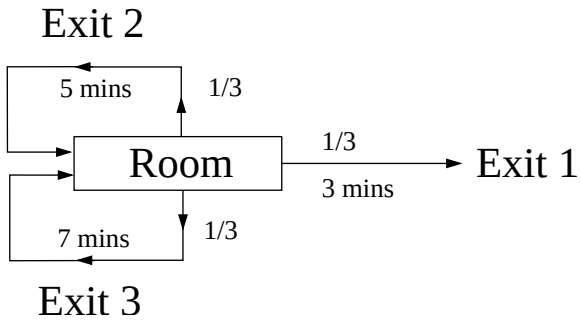
A mouse is trapped in a room with three exits at the centre of a maze.

- Exit 1 leads outside the maze after 3 minutes.
- Exit 2 leads back to the room after 5 minutes.
- Exit 3 leads back to the room after 7 minutes.

Every time the mouse makes a choice, it is equally likely to choose any of the three exits. What is the expected time taken for the mouse to leave the maze?

Let X = time taken for mouse to leave maze, starting from room R.

Let Y = exit the mouse chooses first (1, 2, or 3).



Then:

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}_Y\Big(\mathbb{E}(X \mid Y)\Big) \\ &= \sum_{y=1}^{3} \mathbb{E}(X \mid Y = y)\mathbb{P}(Y = y) \\ &= \mathbb{E}(X \mid Y = 1) \times \tfrac{1}{3} + \mathbb{E}(X \mid Y = 2) \times \tfrac{1}{3} + \mathbb{E}(X \mid Y = 3) \times \tfrac{1}{3}\,.\end{aligned}$$

But:

$$\begin{aligned}\mathbb{E}(X \mid Y = 1) &= 3 \textit{ minutes} \\ \mathbb{E}(X \mid Y = 2) &= 5 + \mathbb{E}(X) \textit{ (after 5 mins back in Room, time } \mathbb{E}(X) \textit{ to get out)} \\ \mathbb{E}(X \mid Y = 3) &= 7 + \mathbb{E}(X) \textit{ (after 7 mins, back in Room)}\end{aligned}$$

So

$$\begin{aligned}\mathbb{E}(X) &= 3 \times \tfrac{1}{3} + \Big(5 + \mathbb{E}X\Big) \times \tfrac{1}{3} + \Big(7 + \mathbb{E}X\Big) \times \tfrac{1}{3} \\ &= 15 \times \tfrac{1}{3} + 2(\mathbb{E}X) \times \tfrac{1}{3} \\ \tfrac{1}{3}\,\mathbb{E}(X) &= 15 \times \tfrac{1}{3} \\ \Rightarrow \quad \mathbb{E}(X) &= 15 \textit{ minutes.}\end{aligned}$$

Notation for quick solutions of first-step analysis problems

As for probabilities, first-step analysis for expectations relies on a good notation. The best way to tackle the problem above is as follows.

Define $m_R = \mathbb{E}(\textit{time to leave maze} \mid \textit{start in Room})$.

First-step analysis:

$$\begin{aligned} m_R &= \tfrac{1}{3} \times 3 + \tfrac{1}{3} \times (5 + m_R) + \tfrac{1}{3} \times (7 + m_R) \\ \Rightarrow \quad 3m_R &= (3 + 5 + 7) + 2m_R \\ \Rightarrow \quad m_R &= 15 \textit{ minutes} \qquad \textit{(as before)}.\end{aligned}$$

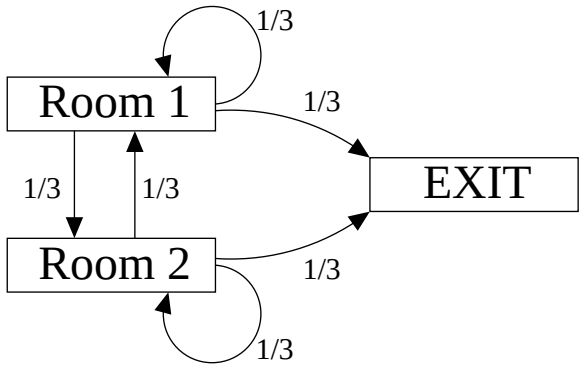
Example 2: Counting the steps

The most common questions involving first-step analysis for expectations ask for the ***expected number of steps before finishing.*** The number of steps is usually equal to the *number of arrows traversed from the current state to the end.*

The key point to remember is that when we take expectations, we are usually *counting something.*

You must remember to *add on whatever we are counting, to every step taken.*

The mouse is put in a new maze with two rooms, pictured here. Starting from Room 1, what is the expected number of steps the mouse takes before it reaches the exit?



1. Define notation: let

$$m_1 = \mathbb{E}(\textit{number of steps to finish} \mid \textit{start in Room 1})$$
$$m_2 = \mathbb{E}(\textit{number of steps to finish} \mid \textit{start in Room 2}).$$

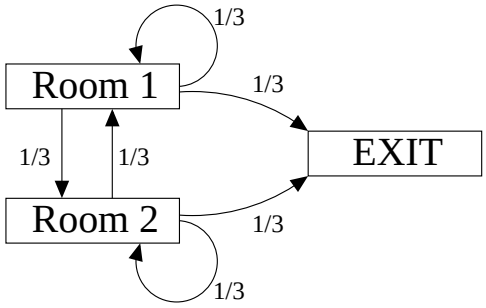
2. First-step analysis:

$$m_1 = \tfrac{1}{3} \times 1 + \tfrac{1}{3}(1 + m_1) + \tfrac{1}{3}(1 + m_2) \qquad (a)$$
$$m_2 = \tfrac{1}{3} \times 1 + \tfrac{1}{3}(1 + m_1) + \tfrac{1}{3}(1 + m_2) \qquad (b)$$

We could solve as simultaneous equations, as usual, but in this case inspection of (a) and (b) shows immediately that $m_1 = m_2$*. Thus:*

$$\textit{(a)} \quad \Rightarrow \quad 3m_1 = 3 + 2m_1$$
$$\Rightarrow \quad m_1 = 3 \textit{ steps.}$$
$$\textit{Further,} \qquad m_2 = m_1 = 3 \textit{ steps also.}$$

Incrementing before partitioning



In many problems, all possible first-step options incur the same initial penalty. The last example is such a case, because *every possible step adds 1 to the total number of steps taken.*

In a case where all steps incur the same penalty, there are two ways of proceeding:

1. *Add the penalty onto each option separately: e.g.*

$$m_1 = \tfrac{1}{3} \times 1 + \tfrac{1}{3}\,(1 + m_1) + \tfrac{1}{3}\,(1 + m_2)\,.$$

2. *(Usually quicker) Add the penalty once only, at the beginning:*

$$m_1 = 1 \quad + \quad \tfrac{1}{3} \times 0 + \tfrac{1}{3} m_1 + \tfrac{1}{3} m_2.$$

In each case, we will get the same answer (check). This is because the option probabilities sum to 1, *so in Method 1 we are adding* $(\frac{1}{3} + \frac{1}{3} + \frac{1}{3}) \times 1 = 1 \times 1 = 1$*, just as we are in Method 2.*

3.6 Probability as a conditional expectation

Recall from Section 3.1 that for any event A, we can write $\mathbb{P}(A)$ as an expectation as follows.

Define the indicator random variable: $I_A = \begin{cases} 1 & \text{if event } A \text{ occurs,} \\ 0 & \text{otherwise.} \end{cases}$

Then $\mathbb{E}(I_A) = \mathbb{P}(I_A = 1) = \mathbb{P}(A)$.

We can refine this expression further, using the idea of conditional expectation. Let Y be any random variable. Then

$$\mathbb{P}(A) = \mathbb{E}(I_A) = \mathbb{E}_Y\Big(\mathbb{E}(I_A \,|\, Y)\Big).$$

But

$$\begin{aligned}\mathbb{E}(I_A \,|\, Y) &= \sum_{r=0}^{1} r\mathbb{P}(I_A = r \,|\, Y) \\ &= 0 \times \mathbb{P}(I_A = 0 \,|\, Y) + 1 \times \mathbb{P}(I_A = 1 \,|\, Y) \\ &= \mathbb{P}(I_A = 1 \,|\, Y) \\ &= \mathbb{P}(A \,|\, Y).\end{aligned}$$

Thus

$$\mathbb{P}(A) = \mathbb{E}_Y\Big(\mathbb{E}(I_A \,|\, Y)\Big) = \mathbb{E}_Y\Big(\mathbb{P}(A \,|\, Y)\Big).$$

This means that for ***any*** random variable X (discrete or continuous), and for any set of values S (a discrete set or a continuous set), we can write:

- for any ***discrete*** random variable Y,

$$\mathbb{P}(X \in S) = \sum_y \mathbb{P}(X \in S \,|\, Y = y)\mathbb{P}(Y = y).$$

- for any ***continuous*** random variable Y,

$$\mathbb{P}(X \in S) = \int_y \mathbb{P}(X \in S \,|\, Y = y) f_Y(y)\, dy.$$

Example of probability as a conditional expectation: winning a lottery

Suppose that a million people have bought tickets for the weekly lottery draw. Each person has a probability of one-in-a-million of selecting the winning numbers. If more than one person selects the winning numbers, the winner will be chosen at random from all those with matching numbers.

You watch the lottery draw on TV and your numbers match the winners!! You had a one-in-a-million chance, and there were a million players, so it must be YOU, right?

Not so fast. Before you rush to claim your prize, let's calculate the probability that you really will win. You definitely win if you are the only person with matching numbers, but you can also win if there there are multiple matching tickets and yours is the one selected at random from the matches.

Define Y to be the number of OTHER matching tickets out of the OTHER 1 million tickets sold. (If you are lucky, $Y = 0$ so you have definitely won.)

If there are 1 million tickets and each ticket has a one-in-a-million chance of having the winning numbers, then

$$Y \sim \textit{Poisson}(1) \textit{ approximately.}$$

The relationship $Y \sim$ Poisson(1) arises because of the Poisson approximation to the Binomial distribution.

(a) What is the probability function of Y, $f_Y(y)$?

$$f_Y(y) = \mathbb{P}(Y = y) = \frac{1^y}{y!}e^{-1} = \frac{1}{e \times y!} \qquad \textit{for } y = 0, 1, 2, \ldots.$$

(b) What is the probability that yours is the only matching ticket?

$$\mathbb{P}(\textit{only one matching ticket}) = \mathbb{P}(Y = 0) = \frac{1}{e} = 0.368.$$

(c) The prize is chosen at random from all those who have matching tickets. What is the probability that you win if there are $Y = y$ OTHER matching tickets?

Let W be the event that I win.

$$\mathbb{P}(W \,|\, Y = y) = \frac{1}{y+1}.$$

(d) Overall, what is the probability that you win, given that you have a matching ticket?

$$\begin{aligned}
\mathbb{P}(W) &= \mathbb{E}_Y\left\{\mathbb{P}(W \mid Y = y)\right\} \\
&= \sum_{y=0}^{\infty} \mathbb{P}(W \mid Y = y)\mathbb{P}(Y = y) \\
&= \sum_{y=0}^{\infty} \left(\frac{1}{y+1}\right)\left(\frac{1}{e \times y!}\right) \\
&= \frac{1}{e}\sum_{y=0}^{\infty} \frac{1}{(y+1)y!} \\
&= \frac{1}{e}\sum_{y=0}^{\infty} \frac{1}{(y+1)!} \\
&= \frac{1}{e}\left\{\sum_{y=0}^{\infty} \frac{1}{y!} - \frac{1}{0!}\right\} \\
&= \frac{1}{e}\{e-1\} \\
&= 1 - \frac{1}{e} \\
&= 0.632.
\end{aligned}$$

Disappointing?

3.7 Special process: a model for gene spread

Suppose that a particular gene comes in two variants (alleles): A and B. We might be interested in the case where one of the alleles, say A, is harmful — for example it causes a disease. All animals in the population must have either allele A or allele B. We want to know how long it will take before all animals have the same allele, and whether this allele will be the harmful allele A or the safe allele B. This simple model assumes asexual reproduction. It is very similar to the famous Wright-Fisher model, which is a fundamental model of population genetics.

Assumptions:

1. The population stays at constant size N for all generations.
2. At the end of each generation, the N animals create N offspring and then they immediately die.
3. If there are x parents with allele A, and $N - x$ with allele B, then each offspring gets allele A with probability x/N and allele B with $1 - x/N$.
4. All offspring are independent.

Stochastic process:

The **state** of the process at time t is $X_t =$ *the number of animals with allele A at generation* t.

Each X_t could be *0, 1, 2, . . . , N.* The state space is $\{0, 1, 2, \ldots, N\}$.

Distribution of $[\, X_{t+1} \,|\, X_t \,]$

Suppose that $X_t = x$, so x of the animals at generation t have allele A.

Each of the N offspring will get A with probability $\frac{x}{N}$ and B with probability $1 - \frac{x}{N}$.

Thus the number of offspring at time $t+1$ with allele A is: $X_{t+1} \sim \textit{Binomial}\left(N, \frac{x}{N}\right)$.

We write this as follows:

$$[\, X_{t+1} \,|\, X_t = x \,] \;\sim\; \textit{Binomial}\left(N,\ \frac{x}{N}\right).$$

If

$$[\, X_{t+1} \,|\, X_t = x \,] \;\sim\; \text{Binomial}\left(N,\, \frac{x}{N}\right),$$

then

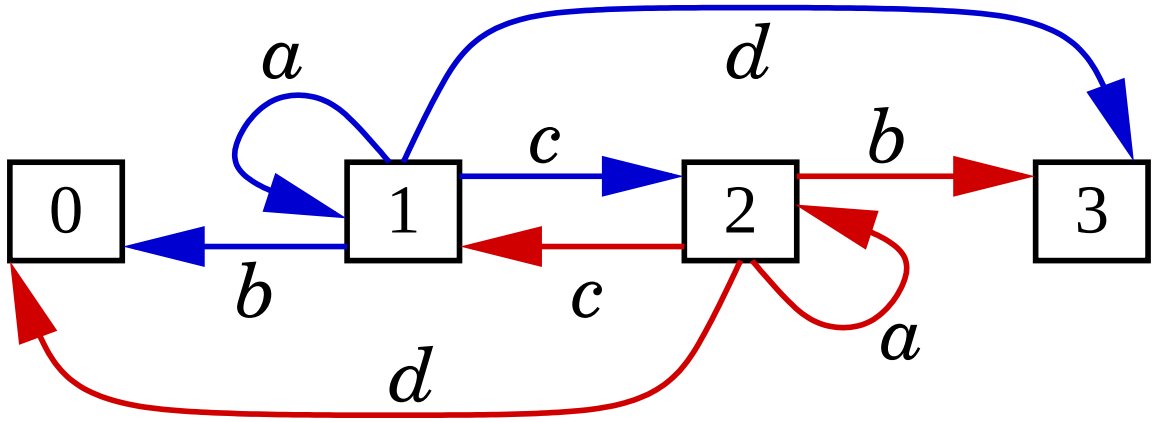
$$\mathbb{P}(X_{t+1} = y \,|\, X_t = x) = \binom{N}{y} \left(\tfrac{x}{N}\right)^y \left(1 - \tfrac{x}{N}\right)^{N-y} \qquad \textit{(Binomial formula)}$$

Example with $N = 3$

This process becomes complicated to do by hand when N is large. We can use small N to see how to use first-step analysis to answer our questions.

Transition diagram:

Exercise: find the missing probabilities a, b, c, and d when $N = 3$. Express them all as fractions over the same denominator.



Probability the harmful allele A dies out

Suppose the process starts at generation 0. One of the three animals has the harmful allele A. Define a suitable notation, and find the probability that the harmful allele A eventually dies out.

Exercise: answer $= 2/3$.

Expected number of generations to fixation

Suppose again that the process starts at generation 0, and one of the three animals has the harmful allele A. Eventually all animals will have the same allele, whether it is allele A or B. When this happens, the population is said to have reached *fixation:* it is fixed for a single allele and no further changes are possible.

Define a suitable notation, and find the expected number of generations to fixation.

Exercise: answer = 3 generations on average.

Things get more interesting for large N. When $N = 100$, and $x = 10$ animals have the harmful allele at generation 0, there is a 90% chance that the harmful allele will die out and a 10% chance that the harmful allele will take over the whole population. The expected number of generations taken to reach fixation is 63.5. If the process starts with just $x = 1$ animal with the harmful allele, there is a 99% chance the harmful allele will die out, but the expected number of generations to fixation is 10.5. Despite the allele being rare, the *average* number of generations for it to either die out or saturate the population is quite large.

Note: The model above is also an example of a process called the ***Voter Process***. The N individuals correspond to N people who each support one of two political candidates, A or B. Every day they make a new decision about whom to support, based on the amount of current support for each candidate. Fixation in the genetic model corresponds to concensus in the Voter Process.