1 Introduction

Here we shall briefly recall the most important facts from basic probability theory; for a more detailed recap, check your notes from the previous years.

1.1 Probability revision

Probability triple

A probability triple is a collection $\big(\Omega, \mathcal{F}, \mathsf{P}\big)$, where:

- Ω is a sample space (eg., the collection $\big\{\mathsf{H}, \mathsf{TH}, \mathsf{TTH}, \dots\big\}$ of all possible results in a coin tossing experiment until the first occurence of H);
- the σ-field $\mathcal{F}$ is the collection of all events under consideration (more precisely it is a set of subsets of Ω satisfying (i) $\varnothing \in \mathcal{F}$ (ii) $A \in \mathcal{F} \Rightarrow \Omega \backslash A \in \mathcal{F}$ & (iii) if $A_k \in \mathcal{F}$ for all $k \geq 1$, then $\cup_{k\geq 1} A_k \in F$); and
- the σ-additive[1] probability measure $\mathsf{P} : \mathcal{F} \to [0,1]$ assigns probabilities to all these events in a consistent way. That is,

$$\mathsf{P} : A \mapsto \mathsf{P}(A) \in [0,1]\,, \qquad \mathsf{P}(\emptyset) = 0\,, \qquad \mathsf{P}(\Omega) = 1, \text{ and}$$

$$(A_k)_{k\geq 1} \in \mathcal{F} \text{ and } A_j \cap A_k = \varnothing \text{ for all } j \neq k \quad \Rightarrow \quad \mathsf{P}(\cup_{k\geq 1} A_k) = \sum_{k\geq 1} \mathsf{P}(A_k).$$

One immediate consequence is monotonicity:

$$C, D \in \mathcal{F} \quad \text{ and } \quad C \subseteq D \qquad \Longrightarrow \qquad \mathsf{P}(C) \leq \mathsf{P}(D)\,.$$

Another important property is the following continuity result:

If the events $A_k \in \mathcal{F}$, $k \geq 1$ form a monotone increasing *sequence, ie., $A_k \subseteq A_{k+1}$ for all $k \geq 1$, then we say that*

$$A_k \nearrow A \overset{\mathsf{def}}{=} \bigcup_{j\geq 1} A_j \text{ as } k \to \infty.$$

We then have that

$$\mathsf{P}(A) = \lim_{k\to\infty} \mathsf{P}(A_k)$$

The proof is a straightforward, but instructive exercise.[2] By taking complements, one deduces an analogous result for monotone decreasing sequences $B_k \in \mathcal{F}$, $k \geq 1$ of events, ie., satisfying $B_k \supseteq B_{k+1}$ for all $k \geq 1$.

Example 1.1. *In the standard coin flipping experiment, let B_k be the event $\{$the first k results are $\mathsf{T}\}$; then $B_k \searrow B \equiv \{$all results are $\mathsf{T}\}$. If the coin shows H with probability $p > 0$, then $\mathsf{P}(B_k) = (1-p)^k \searrow 0$ as $k \to \infty$ and, by continuity, $\mathsf{P}(B) = 0$.*

[1] also called countably additive;

[2] See your 2nd year notes.

Exercise 1.2. *In the setup of Example 1.1, let A_k be the event*

$$\{\textit{no } \mathsf{H} \textit{ observed from } k^{th} \textit{ flip onwards}\}.$$

Show that $\mathsf{P}(A_k) = 0$ for all $k \ge 1$. Verify that $\bigl(A_k\bigr)_{k\ge1}$ is a monotone sequence of events with the limit

$$A \equiv \bigl\{\textit{at most finitely many } \mathsf{H} \textit{ observed}\bigr\}.$$

Use the continuity property of probability measures to deduce that $\mathsf{P}(A) = 0$.

Conditional probability

If A, B are events (ie., A, $B \in \mathcal{F}$) with $\mathsf{P}(B) > 0$, then the conditional probability of A given B is

$$\mathsf{P}\bigl(\,A \mid B\,\bigr) \overset{\mathsf{def}}{=} \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}.$$

Notice that $\mathsf{P}(\,\cdot \mid B\,) : \mathcal{F} \to [0,1]$ is a probability measure.

Formula of total probability (partition theorem)

Events B_1, B_2, $\ldots \in \mathcal{F}$ are said to form a partition of Ω, if they are pairwise disjoint:

$$B_i \cap B_j = \varnothing \text{ for } i \neq j,$$

and cover the whole Ω, ie.,

$$\cup_k B_k = \Omega.$$

If $\bigl\{B_1, B_2, \ldots\bigr\}$ form a partition of Ω, then for every $A \in \mathcal{F}$ the following formula of total probability holds (tacitly assuming that $\mathsf{P}\bigl(A \mid B_k\bigr)\,\mathsf{P}(B_k) \equiv \mathsf{P}(A \cap B_k) = 0$ for $\mathsf{P}(B_k) = 0$):

$$\mathsf{P}(A) = \sum_{k\ge1} \mathsf{P}\bigl(A \mid B_k\bigr)\,\mathsf{P}(B_k).$$

Example 1.3. (Discrete Renewals) *Consider a sequence of events that can only occur at discrete times $k = 1, 2, \ldots$ (eg., a light bulb burns out and is immediately replaced with a new one). Assume that the intervals X between consecutive events have a common distribution $f_k = \mathsf{P}(X = k)$, $k \ge 1$. Let r_n denote the probability that an event occurs at time n; ie., $r_n = \mathsf{P}(A_n)$ with $A_n = \bigl\{$ replacement at time n $\bigr\}$; we shall also assume that a replacement also occurs at time 0 so that $r_0 = 1$. Since the events $B_k = \{$ first bulb burns at time k $\}$ form a countable partition of Ω, and $\mathsf{P}(\,A_n \mid B_k\,) = r_{n-k}$ for all $n \ge k$ (with $r_1 = f_1$), the partition theorem implies*

$$r_n = \sum_{k=1}^{n} r_{n-k}\, f_k\,. \tag{1.1}$$

As $f_0 = 0$, the RHS above is a convolution *of the sequences $(r_n)_{n\geq 0}$ and $(f_k)_{k\geq 0}$; the large n behaviour of r_n can be analysed by taking various transforms (eg., generating functions) of these sequences.*

Random variables

Random variables are “nice” functions $X : \Omega \to \mathbb{R}$ characterized by the fact that for every $a \in \mathbb{R}$ we have $\bigl\{\omega \in \Omega : X(\omega) \leq a\bigr\} \in \mathcal{F}$. In other words, every inverse image $X^{-1}\bigl((-\infty, a]\bigr)$ is an event.

For discrete random variables, ie., those attaining at most countably many values, it is often convenient to replace the previous condition with its equivalent: for every $a \in \mathbb{R}$ the set $X^{-1}\bigl(\{a\}\bigr)$ is an event.

Expectation

If X is a discrete random variable taking values in $\bigl\{x_1, x_2, \dots\bigr\}$ with probabilities

$$p_k = \mathsf{P}(X = x_k) \equiv \mathsf{P}\bigl(\{\omega \in \Omega : X(\omega) = x_k\}\bigr)\,,$$

then if $\sum_{k\geq 1} |x_k|\, p_k < \infty$, we say that X is integrable and the expectation $\mathsf{E}(X)$ of X is given by

$$\mathsf{E}(X) \overset{\mathsf{def}}{=} \sum_{k\geq 1} x_k\, p_k \equiv \sum_{k\geq 1} x_k\, \mathsf{P}(X = x_k)\,.$$

(If $\sum |x_k| p_k = \infty$ we also write $\mathsf{E}(|X|) = \infty$).

It is clear from this definition that if X, Y are two random variables defined on the same probability space with $\mathsf{P}(X \leq Y) = 1$ then $\mathsf{E}(X) \leq \mathsf{E}(Y)$. In particular:

Example 1.4. (Markov’s inequality) *Suppose that X is a random variable taking real values, and $a > 0$. Then*

$$\mathsf{E}(|X|) \leq \mathsf{E}(|X| 1_{X\geq a})) \leq a\mathsf{P}(X \geq a)$$

which rearranges to give

$$\mathsf{P}(|X| \geq a) \leq \frac{\mathsf{E}(|X|)}{a}. \tag{1.2}$$

This is a very useful tool for bounding the tail probabilities *of X (i.e. the LHS above with a large).*

It is also straightforward to verify that expectation is linear: if $X_1, \dots, X_n$ are integrable random variables and $a_1, \dots, a_n \in \mathbb{R}$ then

$$\mathsf{E}(\sum_i a_i X_i) = \sum_i a_i \mathsf{E}(X_i).$$

The conditional expectation of X given $B \in \mathcal{F}$ with $\mathsf{P}(B) > 0$ is computed similarly:

$$\mathsf{E}\big(X \mid B \big) = \sum_{k \geq 1} x_k \, \mathsf{P}(X = x_k \mid B) \, .$$

We then have the following result.

Partition theorem for expectations

If $\big\{ B_1, B_2, \dots \big\}$ form a partition of Ω, then for every random variable X

$$\mathsf{E}(X) = \sum_{k \geq 1} \mathsf{E}\big(X \mid B_k \big) \, \mathsf{P}(B_k) \, .$$

In some cases, the RHS above might not be well defined (i.e., the partial sums of the above series may not converge). However, everything is fine if

$$\sum_{k \geq 1} \big| \mathsf{E}\big(X \mid B_k \big) \big| \, \mathsf{P}(B_k) < \infty$$

(and we have equality in the sense that both sides are equal to ∞ if $\mathsf{P}(X \geq 0) = 1$).

Recall also that for a random variable X with $\mathsf{E}(X^2) < \infty$, we define its variance

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

This measures, roughly speaking, how much the random variable deviates from its expected value. For example, if $\mathsf{P}(X = x) = 1$ for some fixed value x, then $\mathsf{Var}(X) = x^2 - (x)^2 = 0$.

Independence

Events A, $B \in \mathcal{F}$ are independent if

$$\mathsf{P}(A \cap B) = \mathsf{P}(A) \, \mathsf{P}(B) \, .$$

Notice that if A, $B \in \mathcal{F}$ are independent and if $\mathsf{P}(A \mid B)$ is well defined, then $\mathsf{P}(A \mid B) = \mathsf{P}(A)$. Also, if $B \in \mathcal{F}$ satisfies $\mathsf{P}(B) \in \{0, 1\}$, then for every $A \in \mathcal{F}$ the events A and B are independent.

In general, a collection of events $\big(A_\alpha \big)_{\alpha \in \mathcal{A}}$ is independent if every finite subcollection is independent, ie., for every $k \geq 1$ and all $\alpha_1, \dots, \alpha_k \in \mathcal{A}$,

$$\mathsf{P}\big(A_{\alpha_1} \cap A_{\alpha_2} \cap \ldots \cap A_{\alpha_k} \big) = \mathsf{P}(A_{\alpha_1}) \, \mathsf{P}(A_{\alpha_2}) \, \ldots \, \mathsf{P}(A_{\alpha_k}) \, .$$

Two (discrete) random variables, X and Y are independent, if for all x, y,

$$\mathsf{P}\big(X = x, Y = y \big) = \mathsf{P}(X = x) \, \mathsf{P}(Y = y).$$

For general random variables (taking potentially uncountably many values in $\mathbb{R}$), the last condition needs to be replaced with, say,

$$\mathsf{P}\big(X \in [a,b], Y \in [c,d]\big) = \mathsf{P}\big(X \in [a,b]\big)\,\mathsf{P}\big(Y \in [c,d]\big)$$

for all finite or infinite real a, b, c, d. Recall that in the discrete case the collection of numbers $\mathsf{P}\big(X = x, Y = y\big)$ is the joint distribution of the pair (X,Y) of random variables.

Of course, the above idea can also be used to define independence of arbitrary collections of random variables.

Exercise 1.5. *Let D be the result of a single roll of a standard fair dice. Next, flip a fair coin D times, and let H be the total number of* 'heads' *observed. Write the joint distribution of the pair (D,H). Are D and H independent?*

Example 1.6. *If X and Y are* independent discrete random variables*, and f, $g : \mathbb{R} \to \mathbb{R}$ are* arbitrary *functions, then $f(X)$ and $g(Y)$ are* independent random variables *and* $\mathsf{E}\big[f(X)g(Y)\big] = \mathsf{E}f(X) \cdot \mathsf{E}g(Y)$. *For example,* $\mathsf{E}(s^{X+Y}) = \mathsf{E}(s^X)\,\mathsf{E}(s^Y)$ *for* $|s| \le 1$.

Notice that knowing the joint distribution of a random vector (X,Y), we can derive the so-called marginal distributions of its components X and Y. The inverse operation of constructing the joint distribution of a vector from its marginal distributions is not well posed, and often has no unique answer (see below).

Example 1.7. *Let $X \sim \mathsf{Ber}_{\pm1}(p_1)$, ie.,* $\mathsf{P}(X = 1) = 1 - \mathsf{P}(X = -1) = p_1$ *and let Y be a $\mathsf{Ber}_{\pm1}(p_2)$ random variable, ie.,* $\mathsf{P}(Y = 1) = 1 - \mathsf{P}(Y = -1) = p_2$. *Without loss of generality we may assume that $p_1 \le p_2$. Then both of the below are valid joint distributions with given marginals (write $q_i = 1 - p_i$, $i = 1, 2$):*

(A)

	-1	1	X
-1	q_1q_2	q_1p_2	q_1
1	p_1q_2	p_1p_2	p_1
Y	q_2	p_2	

(B)

	-1	1	X
-1	q_2	$p_2 - p_1$	q_1
1	0	p_1	p_1
Y	q_2	p_2	

In (A) the variables X and Y are independent, whereas in (B) they are not independent. This demonstrates that we cannot re-construct the joint distribution of X and Y knowing only their individual distributions.

As we shall see below, this flexibility in constructing several variables on a common probability space (or "coupling") often allows for intuitive and clear probabilistic arguments.

Example 1.8. *Let $(w_n^X)_{n\geq 0}$ be the random walk generated by independent copies of $X \sim \mathsf{Ber}_{\pm 1}(p_1)$; $w_0^X = 0$. Similarly, let $(w_n^Y)_{n\geq 0}$ be the random walk generated by independent copies of $Y \sim \mathsf{Ber}_{\pm 1}(p_2)$; $w_0^Y = 0$. If $p_1 < p_2$, then the law of large numbers implies that $\frac{1}{n}\, w_n^X$ grows linearly with slope $2p_1 - 1$, whereas $\frac{1}{n}\, w_n^Y$ grows linearly with slope $2p_2 - 1 > 2p_1 - 1$. In other words, for times n large enough the trajectories of $(w_n^X)_{n\geq 0}$ will lie below those of $(w_n^Y)_{n\geq 0}$. In fact, using the joint distribution from Example 1.7 (B), one can construct a joint distribution for the entire trajectories of these random walks such that the inequality $w_n^X \leq w_n^Y$ holds for* all *times $n \geq 0$, and not only for n large enough. This is useful. For example, as a result, for* every *monotone increasing function $f : \mathbb{R} \to \mathbb{R}$ one has*

$$f\bigl(w_n^X\bigr) \leq f\bigl(w_n^Y\bigr) \textit{ for all } n \geq 0, \textit{ and therefore } \mathsf{E} f\bigl(w_n^X\bigr) \leq \mathsf{E} f\bigl(w_n^Y\bigr).$$

1.2 Generating functions: key properties

Lengthy calculations arising from even quite straightforward counting problems can be simplified by using generating functions. Recall that the generating function of a real sequence $(a_k)_{k\geq 0}$ is

$$G(s) = G_a(s) \overset{\mathsf{def}}{=} \sum_{k=0}^{\infty} a_k\, s^k \tag{1.3}$$

(defined whenever the sum on the RHS converges). Similarly, the (probability) generating function of a random variable X with values in

$$\mathbb{Z}^+ \overset{\mathsf{def}}{=} \{0, 1, \dots\}$$

is just the generating function of its probability mass function:

$$G(s) \equiv G_X(s) \overset{\mathsf{def}}{=} \mathsf{E}\bigl(s^X\bigr) = \sum_{k=0}^{\infty} s^k \mathsf{P}(X = k)\,. \tag{1.4}$$

Notice that each probability generating function satisfies

$$|G_X(s)| \leq G_X(1) = \sum_{k=0}^{\infty} \mathsf{P}(X = k) \leq 1\,,$$

i.e., is well defined and finite for all (complex) s with $|s| \leq 1$. In particular, $G_X(s)$ can be differentiated term-by-term any number of times in the open unit disk $|s| < 1$.

Generating functions are very useful when studying sums of independent random variables. Indeed, Example 1.6 implies the following important fact:

Example 1.9. *If X and Y are independent random variables with values in $\mathbb{Z}^+$ and $Z = X + Y$, then their generating functions satisfy*

$$G_Z(s) = G_{X+Y}(s) = G_X(s)\, G_Y(s)$$

for all s such that the RHS is well defined.

Example 1.10. *Let $X, X_1, \dots, X_n$ be independent identically distributed random variables*[3] *with values in $\{0, 1, 2, \dots\}$ and let $S_n = X_1 + \dots + X_n$. Suppose that $G_X(s)$ is well-defined. Then*

$$G_{S_n}(s) = G_{X_1}(s) \dots G_{X_n}(s) \equiv \big[G_X(s)\big]^n .$$

Example 1.11. *Let $X, X_1, X_2, \dots$ be i.i.d. with values in $\{0, 1, 2, \dots\}$ and let $N \geq 0$ be an integer-valued random variable independent of $\{X_k\}_{k\geq 1}$. Then $S_N = X_1 + \dots + X_N$ has generating function*

$$G_{S_N} = G_N \circ G_X \tag{1.5}$$

This is a straightforward application of the partition theorem for expectations. Alternatively, *the result follows from the standard properties of conditional expectation:*

$$\mathsf{E}\big(s^{S_N}\big) = \mathsf{E}\Big(\mathsf{E}\big(s^{S_N} \mid N\big)\Big) = \mathsf{E}\Big(\big[G_X(s)\big]^N\Big) = G_N\big(G_X(s)\big) .$$

In general, we say a sequence $\mathbf{c} = (c_n)_{n\geq 0}$ is the convolution of $\mathbf{a} = (a_k)_{k\geq 0}$ and $\mathbf{b} = (b_m)_{m\geq 0}$ (write $\mathbf{c} = \mathbf{a} \star \mathbf{b}$), if[4]

$$c_n = \sum_{k=0}^{n} a_k \, b_{n-k} \, , \qquad n \geq 0 . \tag{1.6}$$

Exercise 1.12. *If $c = a \star b$, show that the generating functions G_c, G_a, and G_b satisfy $G_c = G_a \times G_b$.*

Exercise 1.13. *In the setup of Example 1.3, let G_f and G_r be the generating functions of the sequences $\mathbf{f} = (f_k)_{k\geq 1}$ and $\mathbf{r} = (r_n)_{n\geq 0}$. Show that $G_r(s) = 1/(1 - G_f(s))$ for all $|s| \leq 1$.*

Why do we care? If the generating function G_a of $(a_n)_{n\geq 0}$ is analytic in a neighbourhood of the origin, then there is a one-to-one correspondence between G_a and $(a_n)_{n\geq 0}$. Namely, a_k can be recovered from G_a via[5]

$$a_k = \frac{1}{k!} \frac{d^k}{ds^k} G_a(s) \big|_{s=0} \qquad \text{or} \qquad a_k = \frac{1}{2\pi i} \oint_{|z|=\rho} \frac{G_a(z)}{z^{k+1}} \, dz \, , \tag{1.7}$$

for suitable $\rho > 0$. This result is often referred to as the uniqueness property of generating functions.

[3] from now on we shall often abbreviate this to just i.i.d.

[4] If X and Y are independent variables in $\mathbb{Z}_+$ and $Z = X + Y$, their p.m.f.s satisfy this equation.

[5] if a power series $G_a(s)$ is finite for $|s| < r$ with $r > 0$, then it can be differentiated in the disk $|s| < r$; recall that each probability generating function is analytic in the unit disk $|s| < 1$.

Example 1.14. *Let $X \sim \mathsf{Poi}(\lambda)$ and $Y \sim \mathsf{Poi}(\mu)$ be independent. A straightforward computation gives $G_X(s) = e^{\lambda(s-1)}$ for all s so that if $Z = X + Y$, Example 1.9 implies that*

$$G_Z(s) = G_X(s)\, G_Y(s) = e^{\lambda(s-1)}\, e^{\mu(s-1)} \equiv e^{(\lambda+\mu)(s-1)}\,.$$

This means that Z is $\mathsf{Poi}(\lambda + \mu)$ distributed.

A similar argument can be used in the following exercise.

Exercise 1.15. *If $X \sim \mathsf{Bin}(n, p)$ and $Y \sim \mathsf{Bin}(m, p)$ are independent, show that $X + Y \sim \mathsf{Bin}(n + m, p)$.*

Another useful property of probability generating functions is that they can be used to compute moments:

Theorem 1.16. *If X has generating function G_X, then*

$$\mathsf{E}\big(X(X-1)\dots(X-k+1)\big) = G_X^{(k)}(1)$$

where $G^{(k)}(1)$ is the shorthand for $G^{(k)}(1_-) \equiv \lim_{s\uparrow 1} G^{(k)}(s)$, the limiting value of the kth derivative of $G(s)$ at $s = 1$. Since $s^k G^{(k)}(s)$ is increasing in s, the RHS above is either $+\infty$, or finite. In the latter case, $X(X-1)\dots(X-k+1)$ is integrable and the equality above holds. In the former, $X(X-1)\dots(X-k+1)$ is not integrable.

Exercise 1.17. *Prove Theorem 1.16.*

Remark 1.18. *The quantity $\mathsf{E}\big(X(X-1)\dots(X-k+1)\big)$ is called the kth* factorial moment *of X. Notice also that*

$$\mathsf{Var}(X) = G_X''(1) + G_X'(1) - \big(G_X'(1)\big)^2\,. \tag{1.8}$$

Remark 1.19. *Notice that $\lim_{s\nearrow 1} G_X(s) \equiv \lim_{s\nearrow 1} \mathsf{E}(s^X) = \mathsf{P}(X < \infty)$. This allows us to check whether the random variable X is finite, if we do not know this apriori. See Example 1.26 below.*

Remark 1.20. *The fact that a probability generating function is finite at $u = 1$ (or has a finite left derivative there) does not, in general, imply any regularity beyond the unit disk. Indeed, let X be a random variable satisfying*

$$\mathsf{P}(X = k) = \tfrac{1}{k(k+1)} \qquad \text{for all } k \ge 1,$$

and let G_X be its generating function. It is easy to check that $G_X(1) = 1$ while $\mathsf{E}(X) = G_X'(1_-) = \infty$, and thus $|G_X(u)| \le G_X(1) = 1$ if $|u| \le 1$ but $G_X(u) = \infty$ for all $|u| > 1$.

Exercise 1.21. *Let $\mathsf{P}(X = k) = 4/\big(k(k+1)(k+2)\big)$ for $k \ge 1$. Show that the generating function $G_X(u) = \mathsf{E}(u^X)$ satisfies $G_X(1) = 1$, $G_X'(1_-) = 2 < \infty$, but $G_X''(1_-) = \infty$. Notice that in this case $|G_X'(u)| \le 2$ uniformly in $|u| < 1$ while $G_X(u) = \infty$ for all $|u| > 1$.*

Exercise 1.22. *Following the approach of Exercise 1.21 or otherwise, for $m \in \mathbb{N}$ find a generating function G, which is continuous and bounded on the closed unit disk $|u| \le 1$ together with derivatives up to order m, while $G(u) = \infty$ for all $|u| > 1$.*

Exercise 1.23. *Let $S_N = X_1 + \ldots + X_N$ be a random sum of random variables, whose generating function is $G_{S_N}(u) \equiv G_N\bigl(G_X(u)\bigr)$, recall Example 1.11. Use Theorem 1.16 to express $\mathsf{E}\bigl(S_N\bigr)$ and $\mathsf{Var}\bigl(S_N\bigr)$ in terms of $\mathsf{E}(X)$, $\mathsf{E}(N)$, $\mathsf{Var}(X)$ and $\mathsf{Var}(N)$. Check your result for $\mathsf{E}\bigl(S_N\bigr)$ and $\mathsf{Var}\bigl(S_N\bigr)$ by directly applying the partition theorem for expectations.*

Exercise 1.24. *A bird lays N eggs, each being pink with probability p and blue otherwise. Assuming that $N \sim \mathsf{Poi}(\lambda)$, find the distribution of the total number K of pink eggs.*

Exercise 1.25. *Suppose that in a population, each mature individual produces immature offspring according to a probability generating function F.*

(a) Assume that we start with a population of k immature individuals, each of which grows to maturity with probability p and then reproduces, independently of other individuals. Find the probability generating function of the number of immature individuals in the next generation.

(b) Find the probability generating function of the number of mature individuals in the next generation, given that there are k mature individuals in the parent generation.

(c) Show that the distributions in a) and b) above have the same mean, but not necessarily the same variance. You might prefer to first consider the case $k = 1$, and then generalise.

The next example is very important for applications.

Example 1.26. *Let X_k, $k \ge 1$ be i.i.d. with the common distribution*

$$\mathsf{P}(X_k = 1) = p\,, \qquad \mathsf{P}(X_k = -1) = q = 1 - p\,.$$

Define the simple random walk $(S_n)_{n\ge0}$ via $S_0 = 0$ and $S_n = X_1 + \cdots + X_n$ for $n \ge 1$ and let

$$T \overset{\mathsf{def}}{=} \inf\bigl\{n \ge 1 : S_n = 1\bigr\}$$

be the first time this random walk hits 1.

To calculate the generating function G_T, write $p_k = \mathsf{P}(T = k)$, so that $G_T(s) \equiv \mathsf{E}(s^T) = \sum_{k\ge0} s^k\, p_k$. Conditioning on the outcome of the first step, and applying the partition theorem for expectations, we get

$$G_T(s) \equiv \mathsf{E}\bigl(s^T\bigr) = \mathsf{E}\bigl(s^T \mid X_1 = 1\bigr)\, p + \mathsf{E}\bigl(s^T \mid X_1 = -1\bigr)\, q = ps + qs\, \mathsf{E}\bigl(s^{T_2}\bigr)\,,$$

where T_2 is the time of the first visit to state 1 starting from $S_0 = -1$. By partitioning on the time T_1 of the first visit to 0, it follows that

$$\mathsf{P}(T_2 = m) = \sum_{k=1}^{m-1} \mathsf{P}\bigl(T_1 = k, T_2 = m\bigr)\,.$$

where of course, on the event $\{S_0 = -1\}$,

$$\mathsf{P}(T_2 = m \mid S_0 = -1) \equiv \mathsf{P}\bigl(S_1 < 1, \dots, S_{m-1} < 1, S_m = 1 \mid S_0 = -1\bigr)$$

$$\mathsf{P}(T_1 = k \mid S_0 = -1) \equiv \mathsf{P}\bigl(S_1 < 0, \dots, S_{k-1} < 0, S_k = 0 \mid S_0 = -1\bigr).$$

Notice that by translation invariance the last probability is just $\mathsf{P}(T = k \mid S_0 = 0) = p_k$. We also observe that

$$\mathsf{P}\bigl(T_2 = m \mid T_1 = k\bigr) = \mathsf{P}\bigl(\text{ first hit 1 from 0 after } m-k \text{ steps }\bigr) \equiv p_{m-k}\,.$$

The partition theorem now implies that

$$\mathsf{P}(T_2 = m) = \sum_{k=1}^{m-1} \mathsf{P}\bigl(T_2 = m \mid T_1 = k\bigr)\,\mathsf{P}(T_1 = k) \equiv \sum_{k=1}^{m-1} p_k\,p_{m-k} = \sum_{k=0}^{m} p_k\,p_{m-k}\,,$$

ie., $G_{T_2}(s) = \bigl(G_T(s)\bigr)^2$. We deduce that $G_T(s)$ solves the quadratic equation $\varphi = ps + qs\varphi^2$, so that[6]

$$G_T(s) = \frac{1 - \sqrt{1 - 4pqs^2}}{2qs} = \frac{2ps}{1 + \sqrt{1 - 4pqs^2}}\,.$$

Finally this allows us to deduce that

$$\mathsf{P}(T < \infty) \equiv G_T(1) = \frac{1 - |p-q|}{2q} = \begin{cases} 1\,, & p \ge q\,, \\ p/q\,, & p < q\,. \end{cases}$$

In particular, $\mathsf{E}(T) = \infty$ for $p < q$ (because $\mathsf{P}(T = \infty) = (q-p)/q > 0$). For $p \ge q$ we obtain

$$\mathsf{E}(T) \equiv G_T'(1) = \frac{1 - |p-q|}{2q|p-q|} = \begin{cases} \frac{1}{p-q}\,, & p > q\,, \\ \infty\,, & p = q\,, \end{cases}$$

ie., $\mathsf{E}(T) < \infty$ if $p > q$ and $\mathsf{E}(T) = \infty$ otherwise.

Notice that at criticality ($p = q = 1/2$), the variable T is finite with probability 1, but has infinite expectation.

Example 1.27. *In a sequence of independent Bernoulli experiments with success probability $p \in (0,1)$, let D be the first time that two consecutive successful*

[6] by recalling the fact that $G_T(s) \to 0$ as $s \to 0$;

outcomes have occured (i.e., if two successes occured in the first two experiments, then D would be equal to 2).

To find the generating function of D, there are two good methods, and both involve deriving a recursion relation.

***Method 1**: For $n \geq 2$ let us write $d_n = \mathsf{P}(D = n)$, and consider the events*

- $A := \{(n-2)$ *failures followed by 2 successes*$\}$,
- $A_k := \{$*first failure immediately preceeded by a success occurs at experiment* $k\}$ *for* $k = 2, \ldots, n-2$, *and*
- $B = (A \cup \bigcup_{k=2}^{n-2} A_k)^c$.

These form a partition of the probability space, since they are disjoint by construction, and the definition of B means that they cover the whole space. Note that for $\{D = n\}$ to occur, it must be that one of A or $A_2, \ldots, A_{n-2}$ occurs, so that $\mathsf{P}(\{D = n\} \cap B) = 0$. Also it is clear that $A \subset \{D = n\}$ so $\mathsf{P}(\{D = n\} \cap A) = \mathsf{P}(A) = q^{n-2}p^2$. Thus we can write

$$d_n = \mathsf{P}(D = n) = q^{n-2}p^2 + \sum_{k=2}^{n-2} \mathsf{P}(\{D = n\} \cap A_k)$$

and we are left to calculate $\mathsf{P}(\{D = n\} \cap A_k)$ for $k = 2, \ldots, n-2$. For this, we observe that for $\{D = n\} \cap A_k$ to occur, it must be that the first $(k-2)$ experiments are failures, the $(k-1)st$ is a success, the kth is a failure again, and then for the new sequence of experiments starting from the $(k+1)st$, the first time that 2 consecutive successes are seen is $(n-k)$. By independence of the experiments, the probability of this happening is just $q^{k-2}pqd_{n-k}$. Hence we obtain

$$d_n = q^{n-2}p^2 + \sum_{k=2}^{n-2} q^{k-2}pq\, d_{n-k}\,,$$

and a standard method implies that

$$G_D(s) = \frac{p^2s^2}{1-qs} + \frac{pqs^2}{1-qs}\, G_D(s)\,, \qquad \text{or} \qquad G_D(s) = \frac{p^2s^2}{1-qs-pqs^2}\,.$$

A straightforward computation gives $G_D'(1) = \frac{1+p}{p^2}$, so that on average it takes 42 tosses of a standard symmetric dice until the first two consecutive *sixes appear.*

***Method 2:** we derive a recursion relation directly for the generating function, by conditioning on the result of the first experiment. That is, we use the partition theorem for expectation to write*

$$\begin{aligned} \mathsf{E}(s^D) &= \mathsf{E}(s^D \,|\, \textit{failure})\mathsf{P}(\textit{failure}) + \mathsf{E}(s^D \,|\, \textit{success})\mathsf{P}(\textit{success}) \\ &= s\mathsf{E}(s^{D-1} \,|\, \textit{failure})q + s\mathsf{E}(s^{D-1} \,|\, \textit{success})p. \end{aligned}$$

Now, $\mathsf{E}(s^{D-1}\,|\,\text{failure}\,) = \mathsf{E}(s^D)$ since the new sequence of experiments starting from the 2nd have the same distribution as the whole sequence, and observing a failure for the first experiment means we are still just asking for the first time that two consecutive successes are observed in this new sequence. However, this is not the case for $\mathsf{E}(s^{D-1}\,|\,\text{success}\,)$ since if we have already seen one success this could be the first one in a consecutive pair. So for this conditional expectation we will condition again, but now on the result of the 2nd experiment. If it is a success then s^{D-1} will be s with conditional probability one, and if it is a failure then we will be starting from scratch again. Thus we obtain that $\mathsf{E}(s^{D-1}\,|\,\text{success}\,) = ps + qs\mathsf{E}(s^D)$. Putting this all together and writing $\mathsf{E}(s^D) = G_D(s)$ we see that

$$G_D(s) = qsG_D(s) + p^2s^2 + pqs^2G_D(s)$$

and can rearrange to reach the same conclusion as for method 1.

Exercise 1.28. *In a sequence of independent Bernoulli experiments with success probability $p \in (0,1)$, let M be the first time that m consecutive successful outcomes have occured. Using the approach of Example 1.27 or otherwise, find the generating function of M.*

Exercise 1.29. *In the setup of Example 1.27, show that $d_0 = d_1 = 0$, $d_2 = p^2$, $d_3 = qp^2$, and, conditioning on the value of the first outcome, that $d_n = q\,d_{n-1} + pq\,d_{n-2}$ for $n \geq 3$. Use these relations to re-derive the generating function G_D.*

Exercise 1.30. *Use the method of Exercise 1.29, to derive an alternative solution to Exercise 1.28. Compare the resulting expectation to that in Example 1.27.*

Exercise 1.31. *A biased coin showing 'heads' with probability $p \in (0,1)$ is flipped repeatedly. Let C_w be the first time that the word w appears in the observed sequence of results. Find the generating function of C_w and the expectation $\mathsf{E}\big[C_w\big]$ for each of the following words:* HH, HT, TH *and* TT.

Example 1.32. *If $X_n \sim \mathsf{Bin}(n,p)$ with $p = p_n$ satisfying $n \cdot p_n \to \lambda$ as $n \to \infty$, then $G_{X_n}(s) \equiv \big(1 + p_n(s-1)\big)^n \to \exp\{\lambda(s-1)\}$ for every fixed $s \in [0,1]$, so that the distribution of X_n converges to that of $X \sim \mathsf{Poi}(\lambda)$.*

Exercise 1.33. *For each $n \geq 1$ let $Y_n = \sum_{k=1}^n X_k^{(n)}$, where $X_k^{(n)}$ are independent Bernoulli random variables,*

$$p_k^{(n)} \overset{\mathsf{def}}{=} \mathsf{P}(X_k^{(n)} = 1) = 1 - \mathsf{P}(X_k^{(n)} = 0).$$

Assume that

$$\delta^{(n)} \overset{\mathsf{def}}{=} \max_{1 \leq k \leq n} p_k^{(n)} \to 0$$

as $n \to \infty$ and that for a positive constant λ we have

$$\mathsf{E}(Y_n) \equiv \sum_{k=1}^{n} p_k^{(n)} \to \lambda.$$

Using generating functions or otherwise, show that the distribution of Y_n converges to that of a $\mathsf{Poi}(\lambda)$ random variable. This result is known as the law of rare events.

More generally, we have the following continuity result:

Theorem 1.34. *For every fixed n, suppose that the sequence $a_{0,n}$, $a_{1,n}$, $\dots$ is a probability distribution, ie., $a_{k,n} \ge 0$ and $\sum_{k\ge0} a_{k,n} = 1$, and let G_n be the corresponding generating function, $G_n(s) = \sum_{k\ge0} a_{k,n}s^k$ for all s such that the RHS converges. In order that for every fixed k*

$$\lim_{n\to\infty} a_{k,n} = a_k$$

it is necessary and sufficient that $\lim_{n\to\infty} G_n(s) = G_a(s)$ for every fixed $s \in [0, 1)$, where $G_a(s) = \sum_{k\ge0} a_k s^k$, the generating function of the limiting sequence (a_k).

Remark 1.35. *In the probabilistic context, the convergence above:*

$$a_{k,n} \equiv \mathsf{P}(X_n = k) \overset{n\to\infty}{\to} \mathsf{P}(X = k) = a_k \ \text{ for each } k,$$

is known as convergence in distribution.

Why do we care? In applications one often needs to describe the distribution of a random variable, which is obtained as a result of some limiting approach (or approximation). Then Theorem 1.34 can help to simplify the argument. This method is similar to proving the central limit theorem using moment generating functions $\mathsf{E}(\exp\{tX_n\}) \equiv G_n(e^t)$. Notice that $G_n(e^t)$ exists for some $t > 0$ only if the sequence $a_{k,n} \equiv \mathsf{P}(X_n = k)$ decays sufficiently fast, recall Remark 1.20 above.