BASICS IN APPLIED MATHEMATICS
PART III: STOCHASTICS

PETER PFAFFELHUBER

Contents

Date: Version: August 1st, 2026.

1. Introduction

Stochastics deals with random events. At first it may be surprising that one can find rules governing randomness. Even if one does not know exactly *which* random event occurs, one can often say *with which probability* it occurs. The central notion here is that of a *probability distribution*, which tells us with which (deterministic) probabilities random events happen. We shall not be concerned with how these probabilities are fixed, but mainly with how to compute with them. A second central notion is that of a *random variable*, which describes a random element of a set. In this first section we introduce these notions and, along the way, get to know a few classical models, such as the tossing of coins. These will be very helpful when dealing with concrete (probability) distributions.

In applied mathematics, stochastic methods are needed in many respects. We discuss, for instance, the fundamental area of combinatorics, which is also important for data structures and complexity analysis. Moreover, there is the average-case analysis of algorithms, in which one always assumes a random input. A related question is the analysis of randomised algorithms, i.e. methods in which random decisions are made (although the input is deterministic). Finally, statistical questions play a growing role in computer science, in particular when they are computationally intensive; here we refer to the ever-evolving field of machine learning.

1.1. Frequent models: coin tossing, dice, urns.

Before we formally introduce what probabilities and random variables are, we treat some frequent examples.

Example 1.1 (Coin tossing). When one tosses a coin, it famously shows *heads* or *tails*. For a p-coin toss the probability of *heads* is p, so the probability of *tails* is $q := 1 - p$. For ordinary coins one often thinks of the case $p = 1/2$, because heads and tails come to lie on top with the same probability. If, however, the coin has not the perfect shape, it is clear that $p \neq 1/2$ is also possible. If one tosses the same coin twice, one may ask for the probability of throwing *heads* twice. This is p^2.

Example 1.2 (Dice). A well-known challenge (e.g. in the board game *Mensch ärgere Dich nicht*) is to throw a 6 with a single die as quickly as possible. This is quite similar to the $(1/6)$-coin toss; after all there are only two possibilities, 6 or *not* 6. Thus the probability of throwing no 6 exactly k times is $(5/6)^k$.

Example 1.3 (Random graph). As an application in a different setting, consider a random graph on the vertices $[1 : n] := \{1, \dots, n\}$: for each of the $\binom{n}{2}$ possible edges[1] we toss a p-coin independently and draw the edge if and only if it shows *heads*. This is the Erdős–Rényi model $G(n, p)$. Fix two distinct vertices $u \neq v$ and ask how probable it is that they have a *common neighbour*, i.e. a vertex joined to both; see also Figure 1.1. A third vertex w is a common neighbour precisely when both edges $\{u, w\}$ and $\{v, w\}$ are present. As these are two independent p-coin tosses, this has probability p^2. The $n-2$ candidate vertices w involve pairwise disjoint pairs of edges, so the corresponding events are independent, and the probability that u and v have *no* common neighbour is $(1 - p^2)^{n-2}$. Hence the probability that they have at least one common neighbour is

$$1 - (1 - p^2)^{n-2}.$$

Example 1.4 (Urn). In an urn there are N balls, namely K_1 balls of colour 1, …, K_n balls of colour n (so $K_1 + \cdots + K_n = N$). Drawing (with closed eyes) from the urn, one draws a ball of colour i with probability K_i/N. If one then draws again (without replacing the first ball), the probability of again drawing a ball of colour i is $(K_i - 1)/(N - 1)$. Hence the probability of drawing two balls of the same colour is

$$\sum_{i=1}^{n} \frac{K_i}{N} \frac{K_i - 1}{N - 1}.$$

[1] The possible pairs (u, v) from the set $\{1, ..., n\}$ without replacement, and ignoring the order, are $\{(u, v) \in [1 : n] \mid u < v\}$. There are $\binom{n}{2}$ such pairs.

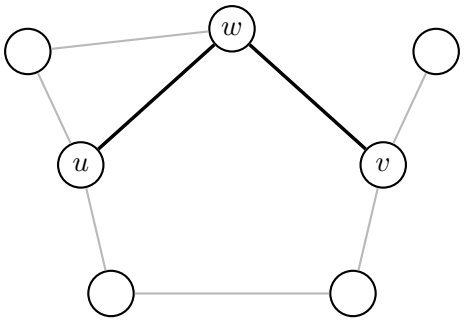

Figure 1.1. In a random graph, the vertex w is a common neighbour of u and v (bold edges): both edges $\{u, w\}$ and $\{v, w\}$ are present.

If, for instance, one has N tasks to be processed by a processor (in random order), namely K_1 tasks of type 1 (e.g. sorting), …, K_n tasks of type n, then this is also the probability that the first two tasks processed are of the same type.

Example 1.5 (Birthday problem). How large is the probability that, among k people, at least two share a birthday? We assume that each person's birthday is uniformly distributed over the $n = 365$ days of the year, and that the k birthdays are independent. It is easier to compute the probability of the complement, that all k birthdays are distinct.

We go through the people in turn. The second person avoids the birthday of the first with probability $(n-1)/n$; the third avoids both earlier birthdays with probability $(n-2)/n$; and so on. Multiplying (the birthdays being independent), the probability that all k are distinct is

$$p_0(k) := \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-k+1}{n} = \prod_{i=1}^{k-1} \Big(1 - \frac{i}{n}\Big),$$

so the probability of at least one shared birthday is $1 - p_0(k)$. For $n = 365$ and $k = 23$ one finds $p_0(23) \approx 0.493$, hence

$$1 - p_0(23) \approx 0.507 > \tfrac{1}{2}.$$

This is the famous *birthday paradox*: already 23 people suffice for a shared birthday to be more likely than not — far fewer than one might guess from the 365 days.

That k can be so much smaller than n is seen from the approximation $1 - x \approx e^{-x}$, valid for small x. It gives

$$p_0(k) \approx \prod_{i=1}^{k-1} e^{-i/n} = e^{-\frac{1}{n}\sum_{i=1}^{k-1} i} = e^{-\frac{k(k-1)}{2n}} \approx e^{-k^2/(2n)}.$$

This is bounded away from both 0 and 1 precisely when the exponent $k^2/(2n)$ is of order 1, i.e. when k is of order $\sqrt{n}$. Hence a collision becomes likely already for $k \approx \sqrt{n}$, not only for $k \approx n$. Solving $e^{-k^2/(2n)} = \frac{1}{2}$ gives the threshold $k \approx \sqrt{2 \ln 2}\,\sqrt{n} \approx 1.18\sqrt{n}$, which for $n = 365$ is about 22.5 — consistent with $k = 23$ above.

1.2. **Random variables and probability distributions.**

We now introduce, still informally, the two central notions.

Definition 1.6 (Random variable). *Let E be a set. A random element of E is called an E-valued)* random variable.

Definition 1.7 (Distribution of a random variable). *Let X be an E-valued random variable.*

(1) *Suppose E is at most countable. Then X is called a* discrete random variable *and the function $x \mapsto \mathbf{P}(X = x)$ is called the* distribution *of X. Here $\mathbf{P}(X = x)$ is the probability that X takes the value x. This map is also called* probability mass function.

(2) *Suppose $E \subseteq \mathbb{R}$.*[2] *If for every interval $A \subseteq \mathbb{R}$*

$$\mathbf{P}(X \in A) = \int_A f(x)\,dx,$$

[2]The resulting continuous distributions will be the topic of Section 4.

then f is called the density *of the distribution of X and X a* continuous *random variable. We then also write $\mathbf{P}(X \in dx) = f(x)\,dx$.*

(3) *We denote by $\mathcal{P}(E)$ the power set of E. Let $\mathbf{P} : \mathcal{A} \subseteq \mathcal{P}(E) \to [0,1]$. If $\mathbf{P}(E) = 1$, $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$ and $\mathbf{P}(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n=1}^{\infty} \mathbf{P}(A_n)$ for pairwise disjoint $A_1, A_2, \dots$ (whenever all values are defined), then $\mathbf{P}$ is called a* probability distribution *on E. If $\mathbf{P}$ and $\mathbf{Q}$ are probability distributions and X is a random variable with $\mathbf{P}(X \in A) = \mathbf{Q}(A)$, we also write $X \sim \mathbf{Q}$.*

Remark 1.8 (Laplace distribution / uniform distribution). The simplest probability distribution on a non-empty finite set E is the one for which all outcomes have the same probability, i.e.

footnoteWe write $|A| := \#A$ for the cardinality of A. If $A := [a,b] \subseteq \mathbb{R}$ is an interval, this might also mean $|A| := b - a$. $\mathbf{P}(A) = \frac{|A|}{|E|}$ for $A \subseteq E$; equivalently $\mathbf{P}(X = a) = \frac{1}{|E|}$ for all $a \in E$. Then X is called (discretely) uniformly distributed on E, and we write $X \sim U(E)$ (Appendix A.8). Such experiments are also called Laplace experiments; examples are the $\frac{1}{2}$-coin toss or throwing a fair die. We return to uniform distributions in Section 1.4.

Example 1.9 (Coin tossing). For a twofold p-coin toss let $X = (X_1, X_2)$ be the random, {heads, tails}-valued vector describing the outcome. Then, for instance,

$$\mathbf{P}(X = (\text{heads, heads})) = \mathbf{P}(X_1 = \text{heads}, X_2 = \text{heads}) = p^2.$$

Remark 1.10 (The *mathematical* definition of distributions). In mathematics one defines random variables and distributions somewhat differently. (Note that in our definition of a random variable from Definition 1.6, note that it is unclear what a *random element of a set* should be.) Whereas above the notion of random variable is in the foreground, in mathematical lectures on stochastics one first introduces probability distributions. For a set Ω these are maps $\mathbf{P} : \mathcal{A} \to [0,1]$, where $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is a σ-algebra, i.e. a system of sets with $\emptyset \in \mathcal{A}$, $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$, and $A_1, A_2, \dots \in \mathcal{A} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$. For the map $\mathbf{P}$ one then requires $\mathbf{P}(\emptyset) = 0$, $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$ and $\mathbf{P}(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n=1}^{\infty} \mathbf{P}(A_n)$ for pairwise disjoint $A_1, A_2, \dots$. In this modelling an E-valued random variable is a map $X : \Omega \to E$, and the distribution of X is given by $\mathbf{P}(X \in A) = \mathbf{P}(X^{-1}(A))$. We shall, however, keep the more intuitive notion of random variable in the foreground. No matter how one models, what matters is that the computational rules of Definition 1.7.3 hold.

We now get to know an important formula for computing with probabilities.

Proposition 1.11 (Inclusion–exclusion formula). *Let X be an E-valued random variable and $A_1, \dots, A_n \subseteq E$. Then*

$$\mathbf{P}(X \in A_1 \cup \dots \cup A_n) = \sum_{i=1}^{n} \mathbf{P}(X \in A_i) - \sum_{1\le i<j\le n} \mathbf{P}(X \in A_i \cap A_j) + \dots \pm \mathbf{P}(X \in A_1 \cap \dots \cap A_n).$$

Proof. First let $n = 2$. Clearly $\mathbf{P}(X \in A_2 \setminus (A_1 \cap A_2)) = \mathbf{P}(X \in A_2) - \mathbf{P}(X \in A_1 \cap A_2)$, since $A_1 \cap A_2$ and $A_2 \setminus (A_1 \cap A_2)$ are disjoint. Hence

$$\begin{aligned}\mathbf{P}(X \in A_1 \cup A_2) &= \mathbf{P}(X \in A_1) + \mathbf{P}(X \in A_2 \setminus (A_1 \cap A_2)) \\ &= \mathbf{P}(X \in A_1) + \mathbf{P}(X \in A_2) - \mathbf{P}(X \in A_1 \cap A_2).\end{aligned}$$

The general case follows by induction on n, splitting $A_1 \cup \dots \cup A_{n+1}$ as $(A_1 \cup \dots \cup A_n) \cup A_{n+1}$ and applying the case $n = 2$ together with the induction hypothesis. □

Example 1.12 (Runs). An algorithm is run repeatedly (with random input). Each time it aborts with probability p and returns the correct result with probability $q := 1 - p$. How large is the probability that, over n runs, it aborts at least k times in a row? More mathematically: we toss a coin with success probability p exactly n times; how large is the probability of a *run* of at least k successes?

The inclusion–exclusion formula (Proposition 1.11) is helpful here. Let X be a random vector of length n with entries in $\{0,1\}$ (where 1 denotes a success and 0 a failure), with

$$\mathbf{P}(X=(x_1,\dots,x_n)) = p^{\sum_{i=1}^n x_i} q^{n-\sum_{i=1}^n x_i}.$$

Further let (with $x_0 := 0$ and fixed k)

$$A_i = \{(x_1,\dots,x_n) : x_{i-1}=0,\ x_i=\dots=x_{i+k-1}=1\},$$

so that $\{X\in A_i\}$ is the event that in X a run of length at least k starts at position i. We wish to compute $\mathbf{P}(X\in A)$ for $A := A_1\cup\dots\cup A_{n-k+1}$. Since the event A_i forces the run to start at i, we have for $i<j$

$$\{X\in A_i\cap A_j\} = \emptyset \text{ if } j-i\le k,$$

and hence, for $i<j$,

$$\mathbf{P}(X\in A_i\cap A_j) = \begin{cases} 0, & \text{if } j-i\le k,\\ q^2p^{2k}, & \text{if } j-i>k,\ i>1,\\ q\,p^{2k}, & \text{if } j-i>k,\ i=1.\end{cases}$$

Applying the inclusion–exclusion formula gives

$$\mathbf{P}(X\in A) = p^k + (n-k)qp^k - (n-2k)qp^{2k} - \binom{n-2k}{2}q^2p^{2k} + \binom{n-3k}{2}q^2p^{3k} + \binom{n-3k}{3}q^3p^{3k}\cdots$$

1.3. **Combinatorial formulae.** In all the examples above it becomes clear that in order to compute probabilities one often has to count something. This art is also called combinatorics. We treat a few selected cases; a summary is found in Table 1.1.

Lemma 1.13 (Combinatorics without repetition). *Let $x=(x_1,\dots,x_n)$ be a vector with distinct entries.*

(1) *There are $n! := n\cdot(n-1)\cdots 1$ distinct vectors of length n having the same entries as x.*
(2) *Drawing k entries without replacement, there are exactly $\frac{n!}{(n-k)!} = n\cdot(n-1)\cdots(n-k+1)$ possible outcomes if the order of the drawn entries is taken into account.*
(3) *Not taking the order into account, there are exactly $\binom{n}{k} := \frac{n!}{(n-k)!K!}$ possible outcomes.*

Proof. 1. When building a vector, the first entry has exactly n possibilities; for each of these the second entry then has $n-1$ possibilities, etc. This gives the formula.
2. follows just as 1., except that one stops after the k-th entry.
3. Starting from 2., there are $\frac{n!}{(n-k)!}$ vectors of length k from the entries of x. Of these, $k!$ each contain the same elements; hence, counting k-element subsets, each subset yields exactly $k!$ vectors, and the result follows. □

Lemma 1.14 (Combinatorics with repetition). *Let $x=(x_1,\dots,x_n)$ be a vector.*

(1) *If exactly r entries are distinct, with multiplicities $k_1,\dots,k_r$, then there are exactly $\binom{n}{k_1\cdots k_r} := \frac{n!}{k_1!\cdots k_r!}$ distinct ways to arrange the entries in order.*
(2) *If all entries are distinct and one draws k times with replacement, there are exactly n^k possible outcomes if the order is taken into account.*
(3) *Not taking the order into account, there are exactly $\binom{n+k-1}{k}$ possible outcomes.*

Proof. 1. If all entries were distinguishable there would be $n!$ possibilities; grouping the $k_1!,\dots,k_r!$ orders of identical entries gives the result.
2. is clear.
3. Here one has to think a little more. Take $n=5$, $k=3$, and let $y=(y_1,\dots,y_5)$ with y_i the number of drawn copies of entry i. Every outcome can be encoded by a code such as $\bullet\bullet{*}{*}\bullet{*}{*}$: two copies of the first entry (hence $\bullet\bullet$), none of the second (no $\bullet$ between the $*$'s), one of the third, and none of the fourth and fifth. This is a bijection between the outcomes and the vectors

	Permutation	Variation	Combination
without repetition	$n!$	$\frac{n!}{(n-k)!}$	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$
example	arrangement of n numbers, each occurring only once	draw k of n numbers, each only once, *with* order	draw k of n numbers, each only once, *without* order
with repetition	$\frac{n!}{k_1! \cdots k_r!}$	n^k	$\binom{n+k-1}{k}$
example	arrangement of n numbers, not all distinct	draw k of n numbers, each arbitrarily often, *with* order	draw k of n numbers, each arbitrarily often, *without* order

Table 1.1. Combinatorics is concerned with counting possibilities. One distinguishes counting with and without repetition, and with (permutations and variations) and without (combinations) regard to order.

of length $n+k-1$ with k entries $\bullet$ and $n-1$ entries $*$ (one needs $n-1$ separators to separate n fields). Their number is $\binom{n+k-1}{k}$. □

Remark 1.15 (Permutation). We call any bijective map of a set $\mathcal{S}$ a permutation of $\mathcal{S}$. Lemma 1.13.1 says that the number of permutations of $[1:n] := \{1, \dots, n\}$ is $|[1:n]| = n!$.

Remark 1.16 (How large is $n!$, actually?). Often, one needs a sense of how large $n!$ is. Stirling's formula is helpful here: it states that

$$\lim_{n\to\infty} \frac{n!}{\left(\frac{n}{e}\right)^n \sqrt{2\pi n}} = 1. \tag{1.1}$$

We shall not prove this, but at least see (which usually suffices for complexity statements) that $n! \approx (n/e)^n$. Indeed,

$$\begin{aligned} \log(n!) &= \sum_{i=1}^{n} \log(i) = \int_1^n \log(x)\, dx + O(n) \\ &= x \log x - x\big|_{x=1}^{n} + O(\log n) = n \log(n) - n + O(\log n), \end{aligned}$$

whence $n! = e^{\log(n!)} = O(n)\left(\frac{n}{e}\right)^n$.

Example 1.17 (Passwords). A password with eight characters from `[a-z, A-Z, 0-9]` is to be generated. For each position there are $26+26+10 = 62$ possibilities, hence $62^8 \approx 2.18 \cdot 10^{14}$ possibilities in total.

Example 1.18 (Merging two lists). Let A be a list of k and B a list of n elements. How many ways are there to arrange the elements of list B into list A (elements of both lists being indistinguishable)? Every such possibility can be written as a vector, e.g. $AAABBABAAABBB\cdots$, of length $n+k$, where A means that an element of list A occurs and B that an element of list B occurs. Since there are $\binom{n+k}{k}$ such vectors, this is also the required number.

Example 1.19 (Lotto: drawing without order or replacement). How many ways are there to choose 6 numbers out of 49 (as in the lottery), disregarding order and without replacement? By Lemma 1.13.3 this is

$$\binom{49}{6} = \frac{49 \cdot 48 \cdots 44}{6 \cdot 5 \cdots 1} = 13\,983\,816.$$

Filling in a single ticket, the probability of matching all six drawn numbers is therefore $1/\binom{49}{6} \approx 7.2 \cdot 10^{-8}$. More generally, the number of correct numbers on a single ticket is *hypergeometrically distributed*, $\mathrm{Hyp}(6, 49, 6)$ (Appendix A.4).

1.4. **Uniform distributions and random permutations.**

Example 1.20 (Random permutations). Below, we will analyse some details of the QUICKSORT algorithm with random input (Section 3.6). Therefore, we consider a random permutation Σ, which has the discrete uniform distribution on $\mathcal{S}_n$, the set of permutations of $\{1, \dots, n\}$. If, for instance,

$$A_j = \{\sigma : \sigma(1) = \max(\sigma(1), \dots, \sigma(j))\},$$

then $|A_j| = \binom{n}{j}(n-j)!(j-1)! = \frac{n!}{j}$ (as the first position among $\sigma(1), \dots, \sigma(j)$ is already fixed). Hence

$$\mathbf{P}(\Sigma \in A_j) = \frac{n!}{n!\, j} = \frac{1}{j}.$$

Example 1.21 (Binomial distribution). Let $(X_1, \dots, X_n)$ be the outcome of an n-fold p-coin toss and $q := 1 - p$, i.e. $X = (X_1, \dots, X_n)$ is a $\{0, 1\}^n$-valued random variable with

$$\mathbf{P}(X = (x_1, \dots, x_n)) = p^{\sum_{i=1}^n x_i}(1-p)^{n - \sum_{i=1}^n x_i}.$$

Let $h(x_1, \dots, x_n) = x_1 + \cdots + x_n$. Then $|h^{-1}(k)| = \binom{n}{k}$, so

$$\mathbf{P}\big(h(X) = k\big) = \sum_{(x_1, \dots, x_n) \in h^{-1}(k)} p^k q^{n-k} = \binom{n}{k} p^k q^{n-k}.$$

This is the distribution of the number of successes; a random variable with this distribution is called binomially distributed with parameters n and p, written $\mathrm{Bin}(n, p)$; see the catalogue in Appendix A.2. If, instead of two outcomes, one rolls a die n times and records how often each of the faces $1, 2, \dots, 6$ appears, one obtains the *multinomial distribution* $\mathrm{Mult}(n; p_1, \dots, p_6)$ (Appendix A.3), the several-category generalisation of the binomial.

2. Independence, product spaces and conditional probability

In several of the examples above (Examples 1.1, 1.2, 1.3, 1.4, 1.5, 1.12) we already multiplied probabilities. For instance, we said that for a p-coin toss the probability of two consecutive *heads* is p^2. We justified this by the independence of the coin tosses. We now make this notion precise, and then turn to conditional probabilities.

2.1. **Stochastic independence.**

Definition 2.1 (Independence). *Let $X_1, \dots, X_n$ be random variables. If*

$$\mathbf{P}(X_1 \in A_1, \dots, X_n \in A_n) = \mathbf{P}(X_1 \in A_1) \cdots \mathbf{P}(X_n \in A_n) \tag{2.1}$$

for arbitrary $A_1, \dots, A_n$, then $(X_1, \dots, X_n)$ is called independent.

Remark 2.2 (Independence of discrete and continuous random variables, and of events). One easily checks the following.

If $X_1, \dots, X_n$ are discrete random variables, then $(X_1, \dots, X_n)$ is independent if and only if

$$\mathbf{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbf{P}(X_1 = x_1) \cdots \mathbf{P}(X_n = x_n)$$

for any choice of $x_1, \dots, x_n$.

If $X_1, \dots, X_n$ are continuous random variables whose joint distribution has a density f, then $(X_1, \dots, X_n)$ is independent if and only if there are functions $f_1, \dots, f_n$ with $f(x_1, \dots, x_n) = f_1(x_1) \cdots f_n(x_n)$. In that case

$$f_i(x_i) = \int f(x_1, \dots, x_n)\, dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n.$$

Independence of *events* is a special case: events $A_1, \dots, A_n$ are called *independent* if their indicator variables $1_{A_1}, \dots, 1_{A_n}$ are independent in the sense of (2.1). Since each 1_{A_i} is discrete, the criterion above shows that this is equivalent to

$$\mathbf{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbf{P}(A_{i_1}) \cdots \mathbf{P}(A_{i_k}) \quad \text{for all } 1 \le i_1 < \dots < i_k \le n.$$

Note that the product formula is required for *all* sub-collections, not just for the full one.

Example 2.3 (Independent coin tosses). We have already met examples of independent random variables. For instance, a p-coin toss $(X_1, \dots, X_n)$ is independent.

Example 2.4 (Minimum and maximum). Let X, Y be independent and identically distributed with an arbitrary continuous distribution, i.e. $\mathbf{P}(X \in dx, Y \in dy) = f(x) \cdot f(y)\, dx\, dy$ for some density f. Then X, Y are independent by construction. The random variables $\min(X, Y)$ and $\max(X, Y)$, however, are *not* independent. Indeed, for $a \le b$,

$$\begin{aligned}\mathbf{P}(\min(X,Y) \le a, \max(X,Y) \le b) &= \mathbf{P}(X \le Y, X \le a, Y \le b) + \mathbf{P}(Y \le X, Y \le a, X \le b)\\ &= 2 \int_{-\infty}^{a} \int_{-\infty}^{b} 1_{x \le y} f(x) f(y)\, dy\, dx.\end{aligned}$$

Thus $(\min(X, Y), \max(X, Y))$ has joint density $2 \cdot 1_{x \le y} f(x) f(y)$, which because of the factor $1_{x \le y}$ does not factorise into a function of x times a function of y. This holds for *every* continuous distribution.

Example 2.5 (Pairwise vs. joint independence). Toss a fair coin twice and let $X_1, X_2 \in \{0, 1\}$ be the two outcomes (with 1 for *heads*), so that X_1, X_2 are independent, each uniformly distributed on $\{0, 1\}$. Put $X_3 := 1_{\{X_1 \ne X_2\}}$, the indicator that exactly one of the two tosses shows *heads* (equivalently $X_3 = X_1 \oplus X_2$, addition modulo 2). Then X_3 is again uniformly distributed on $\{0, 1\}$, and each of the three pairs (X_1, X_2), (X_1, X_3), (X_2, X_3) is independent; for instance, for $a, c \in \{0, 1\}$,

$$\mathbf{P}(X_1 = a, X_3 = c) = \mathbf{P}(X_1 = a, X_2 = a \oplus c) = \tfrac{1}{4} = \mathbf{P}(X_1 = a)\, \mathbf{P}(X_3 = c).$$

Thus X_1, X_2, X_3 are *pairwise* independent. They are nevertheless *not* (jointly) independent, since X_3 is a function of X_1, X_2: for example

$$\mathbf{P}(X_1 = 1, X_2 = 1, X_3 = 1) = 0 \ne \tfrac{1}{8} = \mathbf{P}(X_1 = 1)\, \mathbf{P}(X_2 = 1)\, \mathbf{P}(X_3 = 1).$$

Pairwise independence is thus strictly weaker than (joint) independence.

Proposition 2.6 (Independence is preserved under functions). *If $X_1, \dots, X_n$ are independent random variables and $g_1, \dots, g_n$ are (measurable) maps, then $g_1(X_1), \dots, g_n(X_n)$ are independent.*

Proof. For sets $B_1, \dots, B_n$ we have $\{g_i(X_i) \in B_i\} = \{X_i \in g_i^{-1}(B_i)\}$, so by the independence of $X_1, \dots, X_n$,

$$\begin{aligned}\mathbf{P}(g_1(X_1) \in B_1, \dots, g_n(X_n) \in B_n) &= \mathbf{P}(X_1 \in g_1^{-1}(B_1), \dots, X_n \in g_n^{-1}(B_n))\\ &= \mathbf{P}(X_1 \in g_1^{-1}(B_1)) \cdots \mathbf{P}(X_n \in g_n^{-1}(B_n))\\ &= \mathbf{P}(g_1(X_1) \in B_1) \cdots \mathbf{P}(g_n(X_n) \in B_n),\end{aligned}$$

which is the required factorisation. □

2.2. **Product spaces and product experiments.** We model several experiments run "side by side" directly through independent random variables $X_1, \dots, X_n$, where X_i takes values in an at most countable state space E_i. It is instructive to see the underlying construction of a joint probability space, which we carry out in the discrete mathematical form of Remark 1.10.

Example 2.7 (Two experiments, side by side). Two students carry out an experiment at the same time: Student 1 flips a fair coin, described by a $\{H, T\}$-valued random variable X_1 with $\mathbf{P}(X_1 = H) = \mathbf{P}(X_1 = T) = \frac{1}{2}$; Student 2 rolls a fair die, described by a $\{1, \dots, 6\}$-valued random variable X_2 with $\mathbf{P}(X_2 = k) = \frac{1}{6}$ for $k = 1, \dots, 6$. We would like to speak of the probability of the event "Student 1 throws heads *and* Student 2 rolls a 6", that is, of $\mathbf{P}(X_1 = H, X_2 = 6)$. So far, however, X_1 and X_2 live in two separate experiments, and the joint distribution of (X_1, X_2)

— to which this probability refers — has not yet been specified. Since the two students do not influence one another, we should like X_1 and X_2 to be independent. The remedy is to build a joint space carrying both random variables, on which they are indeed independent.

Definition 2.8 (Product space). *Let $(\Omega_1, \mathcal{P}(\Omega_1), \mathbf{P}_1), \dots, (\Omega_n, \mathcal{P}(\Omega_n), \mathbf{P}_n)$ be discrete probability spaces with counting densities $p_1, \dots, p_n$. The* product space $(\Omega, \mathcal{P}(\Omega), \mathbf{P})$ *is the discrete probability space with*

$$\Omega = \Omega_1 \times \cdots \times \Omega_n = \{(\omega_1, \dots, \omega_n) : \omega_i \in \Omega_i \text{ for all } i\}$$

and counting density $p(\omega_1, \dots, \omega_n) = p_1(\omega_1) \cdots p_n(\omega_n)$. The corresponding measure is called the product measure *and denoted $\mathbf{P} = \mathbf{P}_1 \otimes \cdots \otimes \mathbf{P}_n$.*

Theorem 2.9. *Let $(\Omega_1, \mathcal{P}(\Omega_1), \mathbf{P}_1), \dots, (\Omega_n, \mathcal{P}(\Omega_n), \mathbf{P}_n)$ be discrete probability spaces with product space $(\Omega, \mathcal{P}(\Omega), \mathbf{P})$. On Ω define the* coordinate random variables

$$X_i : \Omega \to \Omega_i, \qquad X_i(\omega) := \omega_i, \qquad i = 1, \dots, n.$$

Then each X_i has distribution $\mathbf{P}_i$, i.e. $\mathbf{P}(X_i \in A_i) = \mathbf{P}_i(A_i)$ for $A_i \subseteq \Omega_i$, and $X_1, \dots, X_n$ are stochastically independent. Equivalently,

$$\mathbf{P}(X_1 \in A_1, \dots, X_n \in A_n) = \mathbf{P}_1(A_1) \cdots \mathbf{P}_n(A_n). \tag{2.2}$$

Proof. Since all summands are non-negative, the (possibly infinite) sums may be rearranged. Observing that $\{X_i \in A_i\} = \{\omega \in \Omega : \omega_i \in A_i\}$, we obtain for the i-th marginal

$$\mathbf{P}(X_i \in A_i) = \sum_{(\omega_1,\dots,\omega_n):\, \omega_i \in A_i} p_1(\omega_1) \cdots p_n(\omega_n) = \Big(\prod_{j \neq i} \underbrace{\sum_{\omega_j \in \Omega_j} p_j(\omega_j)}_{=1} \Big) \sum_{\omega_i \in A_i} p_i(\omega_i) = \mathbf{P}_i(A_i).$$

In the same way, for indices $1 \le i_1 < \cdots < i_k \le n$,

$$\begin{aligned} \mathbf{P}(X_{i_1} \in A_{i_1}, \dots, X_{i_k} \in A_{i_k}) &= \Big(\sum_{\omega_{i_1} \in A_{i_1}} p_{i_1}(\omega_{i_1}) \Big) \cdots \Big(\sum_{\omega_{i_k} \in A_{i_k}} p_{i_k}(\omega_{i_k}) \Big) \\ &= \mathbf{P}(X_{i_1} \in A_{i_1}) \cdots \mathbf{P}(X_{i_k} \in A_{i_k}), \end{aligned}$$

so $X_1, \dots, X_n$ are independent. The special case $k = n$ is exactly (2.2). □

Example 2.10 (Two experiments, side by side, continued). We return to Example 2.7. Set $\Omega := \Omega_1 \times \Omega_2$ with $\Omega_1 = \{H, T\}$, $\Omega_2 = \{1, \dots, 6\}$ and $\mathbf{P} = \mathbf{P}_1 \otimes \mathbf{P}_2$, and let X_1, X_2 be the coordinate random variables from Theorem 2.9, so that X_1 describes Student 1's coin and X_2 Student 2's die. By the theorem, X_1 and X_2 are independent, and the sought-after event "Student 1 throws heads, Student 2 rolls a 6" is $\{X_1 = H,\ X_2 = 6\}$; by (2.2),

$$\mathbf{P}(X_1 = H, X_2 = 6) = \mathbf{P}_1(\{H\}) \cdot \mathbf{P}_2(\{6\}) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}.$$

This is the joint distribution of (X_1, X_2) we were looking for: under the product measure the coin toss X_1 is independent of the die roll X_2.

2.3. **Conditional probabilities.**

Definition 2.11 (Conditional probability). *Let X and Y be random variables.*

(1) *The* conditional probability of $\{Y \in B\}$ given $\{X \in A\}$ *is*

$$\mathbf{P}(Y \in B \,|\, X \in A) := \frac{\mathbf{P}(X \in A, Y \in B)}{\mathbf{P}(X \in A)}.$$

If $\mathbf{P}(X \in A) = 0$, we define the right-hand side to be 0.

(2) *If X is discrete, the* conditional distribution of Y given X *is the map $\mathbf{P}(Y \in B \,|\, X) : x \mapsto \mathbf{P}(Y \in B \,|\, X = x)$. If (X, Y) has a joint density $f(x, y)\, dx\, dy$, the conditional distribution of Y given X has density $\frac{f(x,y)\,dy}{\int f(x,z)\,dz}$, i.e.*

$$\mathbf{P}(Y \in B \,|\, X) : x \mapsto \frac{\int_B f(x, y)\, dy}{\int f(x, z)\, dz}.$$

Example 2.12 (p-coin toss conditioned on the number of successes). Let $X = (X_1, \dots, X_n)$ be a p-coin toss and $S = X_1 + \cdots + X_n$. For $x_1 + \cdots + x_n = k$ the conditional distribution of X given S is

$$\mathbf{P}(X = x \,|\, S = k) = \frac{p^k(1-p)^{n-k}}{\binom{n}{k} p^k (1-p)^{n-k}} = \frac{1}{\binom{n}{k}}.$$

Thus, given $S = k$, X is uniformly distributed over all k-element subsets of $\{1, \dots, n\}$.

Example 2.13 (Minimum and maximum of two uniform random variables). Let U, V be independent and $U([0,1])$-distributed (the continuous uniform distribution; see the catalogue in Appendix A.9). The random variables $X = U \wedge V$ and $Y = U \vee V$ have joint density $2 \cdot 1_{0 \le x \le y \le 1}$. Hence the conditional distribution of X given $Y = y$ has density

$$\mathbf{P}(X \in dx \,|\, Y = y) = \frac{2 \cdot 1_{0 \le x \le y \le 1}}{2 \int 1_{0 \le x \le y \le 1}\, dx} = \frac{1}{y} 1_{0 \le x \le y},$$

which is precisely the density of a $U([0, y])$ distribution.

Lemma 2.14 (Simple properties of conditional probabilities). *Let X and Y be random variables and A, B sets.*

(1) *If Y is discrete, then*

$$\mathbf{P}(Y \in B \,|\, X \in A) = \sum_{y \in B} \mathbf{P}(Y = y \,|\, X \in A).$$

(2) *X and Y are independent if and only if $\mathbf{P}(Y \in B \,|\, X \in A) = \mathbf{P}(Y \in B)$ for all A, B; equivalently, $\mathbf{P}(Y \in B \,|\, X) = \mathbf{P}(Y \in B)$ for all B.*

Proof. Both properties follow directly from the definition of conditional probability. □

Theorem 2.15 (Law of total probability and Bayes' formula). *Let X, Y be discrete random variables. Then the law of total probability holds:*

$$\mathbf{P}(Y \in B) = \sum_x \mathbf{P}(Y \in B \,|\, X = x) \cdot \mathbf{P}(X = x)$$

for every set B. Moreover, Bayes' formula holds for every set A with $\mathbf{P}(X \in A) > 0$:

$$\mathbf{P}(X \in A \,|\, Y \in B) = \frac{\mathbf{P}(Y \in B \,|\, X \in A) \cdot \mathbf{P}(X \in A)}{\sum_x \mathbf{P}(Y \in B \,|\, X = x) \cdot \mathbf{P}(X = x)}.$$

Proof. The law of total probability follows by inserting the definition of $\mathbf{P}(Y \in B \,|\, X = x)$ into

$$\mathbf{P}(Y \in B) = \sum_x \mathbf{P}(Y \in B, X = x).$$

In Bayes' formula, the numerator of the right-hand side equals $\mathbf{P}(X \in A, Y \in B)$ by definition of conditional probability, and the denominator is

$$\sum_x \mathbf{P}(Y \in B \,|\, X = x) \cdot \mathbf{P}(X = x) = \sum_x \mathbf{P}(Y \in B, X = x) = \mathbf{P}(Y \in B).$$

□

Example 2.16 (A rigged coin). In an urn there are three coins, namely two fair coins and one unfair coin that shows *heads* with probability $p \neq 1/2$. More precisely, let $\mathbf{P}(X_f = \text{heads}) = \mathbf{P}(X_f = \text{tails}) = \frac{1}{2}$, i.e. X_f is the outcome of a fair coin, and $\mathbf{P}(X_u = \text{heads}) = 1 - \mathbf{P}(X_u = \text{tails}) = p$, i.e. X_u is the outcome of the unfair coin. We want to determine the probability that the unfair coin was tossed, given that the toss showed *heads*; see also the probability tree in Figure 2.1.

Here $I = f$ or $I = u$ according to whether a fair or the unfair coin was drawn in the first step. For $X := X_I$ (i.e. $X = X_u$ with probability $\frac{1}{3}$ and $X = X_f$ with probability $\frac{2}{3}$),

$$\mathbf{P}(I = u \,|\, X = \text{heads}) = \frac{\mathbf{P}(I = u, X = \text{heads})}{\mathbf{P}(I = f, X = \text{heads}) + \mathbf{P}(I = u, X = \text{heads})} = \frac{\frac{1}{3}p}{\frac{1}{3} + \frac{1}{3}p} = \frac{p}{1+p}.$$

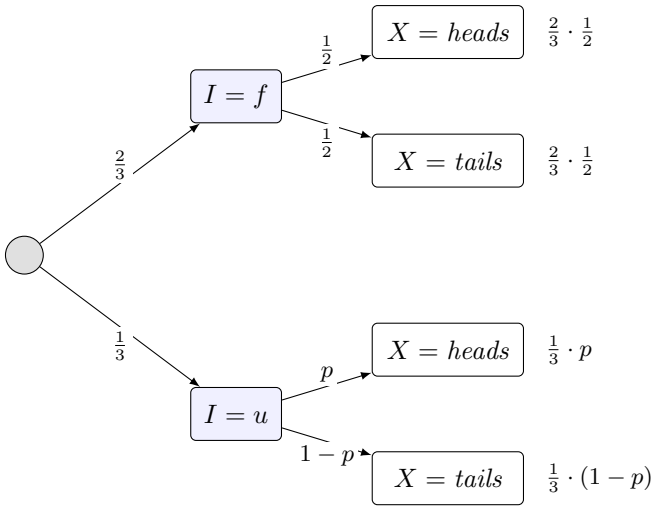

Figure 2.1. The probability tree of Example 2.16. Each path probability is the product of the edge probabilities along the path.

3. Discrete random variables

The distribution of a random variable is often complicated; think, for instance, of the (random) number of comparisons that Quicksort needs to sort a vector. It is therefore often helpful to summarise a distribution by a few characteristic quantities. The most important are certainly the expectation and the variance; for multidimensional distributions the covariance is added. Before we get there, we record the distribution function and one important approximation, the Poisson limit theorem.

3.1. Distribution function and quantiles.

Definition 3.1 (Distribution function and quantiles). *Let $E \subseteq \mathbb{R}$ and X an E-valued random variable. Then $x \mapsto F_X(x) := \mathbf{P}(X \le x)$ is called the* distribution function *of X. For $\alpha \in (0,1)$, any $q_\alpha \in \mathbb{R}$ with $\mathbf{P}(X < q_\alpha) \le \alpha \le \mathbf{P}(X \le q_\alpha)$ is called an* α-quantile *of X. Any* 0.5*-quantile is a* median *of X.*

Lemma 3.2 (Properties of distribution functions). *Let X be a (discrete or continuous) real-valued random variable. Then $\lim_{x\to\infty} F_X(x) = 1$ and $\lim_{x\to-\infty} F_X(x) = 0$; moreover F_X is monotonically increasing. If X is continuous with density $f(x)\,dx$, then*

$$F_X(x) = \int_{-\infty}^{x} f(z)\,dz, \qquad \int_{-\infty}^{\infty} f(z)\,dz = 1, \qquad F_X' = f.$$

Proof. For discrete X with range $E \subseteq \mathbb{R}$ the masses sum to one, hence

$$\mathbf{P}(X \le x) = \sum_{z \in E,\, z \le x} \mathbf{P}(X = z) \xrightarrow{x\to\infty} 1,$$

and the limit at $-\infty$ is analogous. For continuous X, since $(-\infty, x]$ is an interval,

$$\mathbf{P}(X \le x) = \int_{-\infty}^{x} f(z)\,dz.$$

□

Example 3.3 (Waiting time for the first success). Toss a p-coin repeatedly and independently, and let X be the number of tosses up to and including the first head. Then $\{X > k\}$ is the event that the first k tosses all show tails, so $\mathbf{P}(X > k) = (1-p)^k$ for $k = 0, 1, 2, \ldots$; equivalently $\mathbf{P}(X = k) = (1-p)^{k-1}p$, the geometric distribution geo(p) (Appendix A.5). Here the distribution function can be given in closed form: for $x \ge 0$,

$$F_X(x) = \mathbf{P}(X \le x) = 1 - \mathbf{P}(X > \lfloor x \rfloor) = 1 - (1-p)^{\lfloor x \rfloor},$$

a right-continuous step function that vanishes for $x < 1$ and jumps by $\mathbf{P}(X = k)$ at each integer $k \geq 1$. Inverting F_X, the smallest x with $F_X(x) \geq \alpha$ is the α-quantile

$$q_\alpha = \left\lceil \frac{\ln(1-\alpha)}{\ln(1-p)} \right\rceil.$$

For a fair coin ($p = \frac{1}{2}$), for instance, $q_{0.9} = \lceil \ln(0.1)/\ln(0.5) \rceil = \lceil 3.32 \rceil = 4$: with probability at least 90% the first head has appeared within four tosses.

3.2. **Law of small numbers.**

Example 3.4 (Thrice flipping a coin). Let (X_1, X_2, X_3) be a fair coin tossed three times, i.e. independent random variables, each uniformly distributed on $\{0, 1\}$ (with 1 for heads). By independence, $\mathbf{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$ for every $(x_1, x_2, x_3) \in \{0, 1\}^3$. Let $X = X_1 + X_2 + X_3$ count the heads. Summing over the outcomes with exactly k heads, for $k = 0, 1, 2, 3$,

$$\mathbf{P}(X = k) = \sum_{\substack{(x_1,x_2,x_3)\in\{0,1\}^3 \\ x_1+x_2+x_3=k}} \left(\tfrac{1}{2}\right)^3 = \binom{3}{k}\left(\tfrac{1}{2}\right)^3,$$

that is, $X \sim \mathrm{Bin}(3, \frac{1}{2})$, in accordance with the general computation of the binomial distribution above. For instance $\mathbf{P}(X = 1) = \mathbf{P}\big((X_1, X_2, X_3) \in \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}\big) = \frac{3}{8}$. This is the prototypical passage from independent coin tosses to the distribution of a random variable built from them.

Proposition 3.5 (Poisson limit theorem / law of small numbers). *Let $X_n \sim \mathrm{Bin}(n, p_n)$ for $n = 1, 2, \dots$ with $n\,p_n \xrightarrow{n\to\infty} \lambda > 0$ (that is, the expected number of successes np_n converges to λ; expectation is treated in Section 3.3). Then*

$$\lim_{n\to\infty} \mathbf{P}(X_n = k) = e^{-\lambda}\frac{\lambda^k}{k!} \qquad (k = 0, 1, 2, \dots).$$

Proof. We write

$$\begin{aligned}\mathbf{P}(X_n = k) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{n^k} \cdot \frac{1}{k!}\,(np_n)^k \Big(1 - \frac{np_n}{n}\Big)^n (1-p_n)^{-k}.\end{aligned}$$

As $n \to \infty$, $\frac{n(n-1)\cdots(n-k+1)}{n^k} \to 1$, $(1 - np_n/n)^n \to e^{-\lambda}$ and $(1-p_n)^{-k} \to 1$, which gives the claim. □

Remark 3.6 (Poisson distribution). The limit distribution appearing in Proposition 3.5 has a name: a $\mathbb{Z}_+$-valued random variable X with $\mathbf{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ (for a parameter $\lambda \geq 0$) is said to have the *Poisson distribution* with parameter λ, written $X \sim \mathrm{Poi}(\lambda)$. One computes $\mathbf{E}[X] = \mathbf{Var}[X] = \lambda$ (expectation and variance are introduced in Section 3.3; the computation is carried out in Appendix A.7).

3.3. **Expectation and variance.**

Definition 3.7 (Expectation). *Let X be an E-valued random variable and $h : E \to \mathbb{R}$.*

(1) *If E is discrete, then*

$$\mathbf{E}[h(X)] := \sum_{x\in E} h(x)\,\mathbf{P}(X = x)$$

is called the expectation *of $h(X)$, provided the sum converges.*

(2) *If X is continuous with density f, then*

$$\mathbf{E}[h(X)] := \int h(x)\,f(x)\,dx,$$

provided the integral exists.

In particular, for $E \subseteq \mathbb{R}$ and $h = \mathrm{id}$,

$$\mathbf{E}[X] := \sum_{x \in E} x\, \mathbf{P}(X = x) \qquad \textit{resp.} \qquad \mathbf{E}[X] := \int x\, f(x)\, dx,$$

is the expectation of X.

The expectation of the standard distributions — discrete and continuous alike — is computed, distribution by distribution, in the catalogue of Appendix A. For instance, $\mathbf{E}[X] = np$ for the binomial $X \sim \mathrm{Bin}(n, p)$ (Appendix A.2) and $\mathbf{E}[X] = \frac{a+b}{2}$ for the continuous uniform $X \sim U([a, b])$ (Appendix A.9); the corresponding variances are found there as well.

Remark 3.8 (Transformation theorem). Let X be a discrete E-valued random variable and $h : E \to \mathbb{R}$. Then

$$\mathbf{E}[h(X)] = \sum_{y \in h(E)} y\, \mathbf{P}(h(X) = y),$$

provided the sum exists. Indeed,

$$\mathbf{E}[h(X)] = \sum_{x \in E} h(x) \mathbf{P}(X = x) = \sum_{y \in h(E)} y \sum_{x \in h^{-1}(y)} \mathbf{P}(X = x) = \sum_{y \in h(E)} y\, \mathbf{P}(h(X) = y).$$

Example 3.9 (Indicator function). Let X be an E-valued random variable and $1_A : E \to \{0, 1\}$ the indicator of A. Then

$$\mathbf{E}[1_A(X)] = \sum_{y=0,1} y\, \mathbf{P}(1_A(X) = y) = \mathbf{P}(1_A(X) = 1) = \mathbf{P}(X \in A).$$

Lemma 3.10. *Let X be a $\mathbb{Z}_+$-valued random variable. Then, provided the sums exist,*

$$\mathbf{E}[X] = \sum_{x=0}^{\infty} \mathbf{P}(X > x), \qquad \mathbf{E}[X(X-1)] = 2 \sum_{x=0}^{\infty} x\, \mathbf{P}(X > x).$$

If X is $\mathbb{R}_+$-valued with density f, then

$$\mathbf{E}[X] = \int_0^{\infty} \mathbf{P}(X > x)\, dx, \qquad \mathbf{E}[X^2] = 2 \int_0^{\infty} x\, \mathbf{P}(X > x)\, dx.$$

Proof. For the first claim, by absolute convergence we may reorder:

$$\mathbf{E}[X] = \sum_{x=1}^{\infty} \sum_{y=1}^{x} \mathbf{P}(X = x) = \sum_{y=1}^{\infty} \sum_{x=y}^{\infty} \mathbf{P}(X = x) = \sum_{y=0}^{\infty} \mathbf{P}(X > y).$$

The remaining identities are shown analogously. □

These moments are worked out, distribution by distribution, in the catalogue of Appendix A: for the discrete uniform distribution on $\{k, \dots, \ell\}$ one finds $\mathbf{E}[X] = \frac{k+\ell}{2}$ by direct summation (Appendix A.8), and the tail formulas give $\mathbf{E}[X] = \frac{1}{\lambda}$ and $\mathbf{E}[X^2] = \frac{2}{\lambda^2}$ for the exponential distribution $\exp(\lambda)$ (Appendix A.10).

Proposition 3.11 (Properties of the expectation). *Let X, Y be (discrete or continuous) random variables.*

1. *For $a, b \in \mathbb{R}$, $\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y]$, provided all expectations exist.*
2. *If $X \le Y$, then $\mathbf{E}[X] \le \mathbf{E}[Y]$.*

Proof. We treat the discrete case. For 1.,

$$\mathbf{E}[aX + bY] = \sum_{x,y} (ax + by) \mathbf{P}(X = x, Y = y) = a\mathbf{E}[X] + b\mathbf{E}[Y].$$

For 2. it suffices to show $\mathbf{E}[X] \ge 0$ for $X \ge 0$ (then $\mathbf{E}[Y] - \mathbf{E}[X] = \mathbf{E}[Y - X] \ge 0$), which is clear from the definition. □

Example 3.12 (p-coin toss). If $X = (X_1, \dots, X_n)$ is a p-coin toss, then $Y := X_1 + \cdots + X_n \sim \mathrm{Bin}(n, p)$, and by linearity

$$\mathbf{E}[Y] = \mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n] = n\mathbf{P}(X_1 = 1) = np,$$

in agreement with the value $\mathbf{E}[X] = np$ for $\mathrm{Bin}(n, p)$ (Appendix A.2).

Definition 3.13 (Variance). *Let X be a discrete $\mathbb{R}$-valued or a continuous random variable with $\mu := \mathbf{E}[X]$. Then $\sigma^2 := \mathbf{Var}[X] = \mathbf{E}[(X-\mu)^2]$ is the* variance *of X, and $\sigma := \sqrt{\mathbf{Var}[X]}$ its* standard deviation.

Proposition 3.14 (Computing the variance). *Under the above assumptions, $\mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = \mathbf{E}[X(X-1)] + \mathbf{E}[X] - \mathbf{E}[X]^2$, provided all expectations exist.*

Proof. The second equality is clear by linearity. With $\mu := \mathbf{E}[X]$,

$$\mathbf{Var}[X] = \mathbf{E}[X^2 - 2\mu X + \mu^2] = \mathbf{E}[X^2] - \mu^2.$$

□

This formula for the variance is applied, distribution by distribution, in the catalogue of Appendix A: for example it gives $\mathbf{Var}[X] = \frac{(b-a)^2}{12}$ for the continuous uniform distribution $U([a, b])$ (Appendix A.9) and $\mathbf{Var}[X] = npq$ for the binomial distribution $\mathrm{Bin}(n, p)$ (Appendix A.2).

The following two inequalities are fundamental; they turn control of expectation and variance into control of probabilities.

Proposition 3.15 (Markov and Chebyshev inequalities). *Let X be a $\mathbb{R}_+$-valued random variable. Then for all $\varepsilon > 0$ the* Markov inequality $\mathbf{P}(X \geq \varepsilon) \leq \frac{1}{\varepsilon}\mathbf{E}[X]$ *holds. If $\mu := \mathbf{E}[X]$ exists, then for $\varepsilon > 0$ the* Chebyshev inequality $\mathbf{P}(|X - \mu| \geq \varepsilon) \leq \frac{\mathbf{Var}[X]}{\varepsilon^2}$ *holds.*

Proof. Since $X \geq \varepsilon \cdot 1_{X \geq \varepsilon}$, monotonicity of the expectation gives $\varepsilon\,\mathbf{P}(X \geq \varepsilon) = \mathbf{E}[\varepsilon\, 1_{X \geq \varepsilon}] \leq \mathbf{E}[X]$. Chebyshev follows by applying Markov to $(X - \mu)^2$. □

3.4. **Multidimensional distributions.**

Definition 3.16 (Joint distribution). *If $X = (X_1, \dots, X_n)$ has target $E_1 \times \cdots \times E_n$, then the distribution of X is the* joint distribution *of $X_1, \dots, X_n$. The map $A_i \mapsto \mathbf{P}(X_i \in A_i)$ is the i-th* marginal distribution. *If, for all intervals $A_1, \dots, A_n$,*

$$\mathbf{P}(X \in A_1 \times \cdots \times A_n) = \int f(x_1, \dots, x_n)\, dx_1 \cdots dx_n,$$

then f is the density *of the joint distribution.*

Definition 3.17 (Covariance and correlation). *Let X, Y be random variables whose expectations μ_X, μ_Y and variances σ_X^2, σ_Y^2 exist. Then*

$$\mathbf{Cov}[X, Y] := \mathbf{E}[(X - \mu_X)(Y - \mu_Y)]$$

is the covariance *of X and Y; if $\mathbf{Cov}[X, Y] = 0$, then X, Y are called* uncorrelated. *Moreover, $\mathbf{Cor}[X, Y] := \frac{\mathbf{Cov}[X,Y]}{\sigma_X \sigma_Y}$ is the* correlation coefficient.

Lemma 3.18 (Rules for covariances). *Let X, Y, Z be real-valued random variables with finite variance. Then $\mathbf{Cov}[X, Y] = \mathbf{Cov}[Y, X]$, $\mathbf{Var}[X] = \mathbf{Cov}[X, X]$, $\mathbf{Cov}[X, Y] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$, and $\mathbf{Cov}[X, aY + bZ] = a\mathbf{Cov}[X, Y] + b\mathbf{Cov}[X, Z]$.*

Proof. Write $\mu_X = \mathbf{E}[X]$ and $\mu_Y = \mathbf{E}[Y]$. Symmetry is immediate, since the product is commutative:

$$\mathbf{Cov}[X, Y] = \mathbf{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbf{E}[(Y - \mu_Y)(X - \mu_X)] = \mathbf{Cov}[Y, X],$$

and $\mathbf{Cov}[X, X] = \mathbf{E}[(X - \mu_X)^2] = \mathbf{Var}[X]$ by the definition of the variance. For the third identity, expanding the product and using linearity of the expectation,

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

Finally, applying this third identity to $\mathbf{Cov}[X, aY + bZ]$ and then linearity of the expectation,

$$\begin{aligned}\mathbf{Cov}[X, aY + bZ] &= \mathbf{E}[X(aY + bZ)] - \mathbf{E}[X]\,\mathbf{E}[aY + bZ] \\ &= a\big(\mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]\big) + b\big(\mathbf{E}[XZ] - \mathbf{E}[X]\mathbf{E}[Z]\big) \\ &= a\,\mathbf{Cov}[X, Y] + b\,\mathbf{Cov}[X, Z].\end{aligned}$$

□

Proposition 3.19 (Independence implies uncorrelatedness). *If X, Y are independent real-valued random variables with finite variance, then $\mathbf{Cov}[X, Y] = 0$.*

Proof. We treat the discrete case; the continuous case is analogous, with the joint density factorising. By independence, $\mathbf{P}(X = x, Y = y) = \mathbf{P}(X = x)\,\mathbf{P}(Y = y)$, so the double sum factorises into a product of single sums:

$$\begin{aligned}\mathbf{E}[XY] &= \sum_{x,y} xy\,\mathbf{P}(X = x, Y = y) = \sum_{x,y} xy\,\mathbf{P}(X = x)\,\mathbf{P}(Y = y) \\ &= \Big(\sum_x x\,\mathbf{P}(X = x)\Big)\Big(\sum_y y\,\mathbf{P}(Y = y)\Big) = \mathbf{E}[X]\,\mathbf{E}[Y].\end{aligned}$$

Hence, by the third identity of Lemma 3.18, $\mathbf{Cov}[X, Y] = \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y] = 0$. □

Proposition 3.20 (Variance of a sum). *Let $X_1, \dots, X_n$ be real-valued random variables with finite variances. Then*

$$\mathbf{Var}\Big[\sum_{i=1}^n X_i\Big] = \sum_{i=1}^n \mathbf{Var}[X_i] + 2 \sum_{1 \le i < j \le n} \mathbf{Cov}[X_i, X_j].$$

If $(X_1, \dots, X_n)$ is pairwise uncorrelated, then

$$\mathbf{Var}\Big[\sum_i X_i\Big] = \sum_i \mathbf{Var}[X_i].$$

Proof. Using $\mathbf{Var}[Y] = \mathbf{Cov}[Y, Y]$ and then the bilinearity of the covariance (Lemma 3.18, applied in each argument),

$$\mathbf{Var}\Big[\sum_{i=1}^n X_i\Big] = \mathbf{Cov}\Big[\sum_{i=1}^n X_i, \sum_{j=1}^n X_j\Big] = \sum_{i=1}^n \sum_{j=1}^n \mathbf{Cov}[X_i, X_j].$$

Splitting off the diagonal terms $i = j$ and pairing $\mathbf{Cov}[X_i, X_j] = \mathbf{Cov}[X_j, X_i]$ for $i \neq j$ (symmetry),

$$\sum_{i,j} \mathbf{Cov}[X_i, X_j] = \sum_{i=1}^n \mathbf{Cov}[X_i, X_i] + \sum_{i \neq j} \mathbf{Cov}[X_i, X_j] = \sum_{i=1}^n \mathbf{Var}[X_i] + 2 \sum_{1 \le i < j \le n} \mathbf{Cov}[X_i, X_j],$$

which is the first claim. If $(X_1, \dots, X_n)$ is pairwise uncorrelated, every off-diagonal covariance vanishes, leaving

$$\mathbf{Var}\Big[\sum_i X_i\Big] = \sum_i \mathbf{Var}[X_i].$$

□

Example 3.21 (Convolution of binomials). Let $X \sim \mathrm{Bin}(n, p)$ and $Y \sim \mathrm{Bin}(m, p)$ be independent. Conditioning on the value of X (law of total probability) and using independence, we get for $k = 0, \dots, n + m$

$$\begin{aligned}\mathbf{P}(X + Y = k) &= \sum_{l=0}^k \mathbf{P}(X = l)\,\mathbf{P}(Y = k - l) = \sum_{l=0}^k \binom{n}{l} p^l q^{n-l} \binom{m}{k-l} p^{k-l} q^{m-(k-l)} \\ &= p^k q^{n+m-k} \sum_{l=0}^k \binom{n}{l}\binom{m}{k-l} = \binom{n+m}{k} p^k q^{n+m-k},\end{aligned}$$

the last step being the Vandermonde identity $\sum_{l=0}^{k}\binom{n}{l}\binom{m}{k-l}=\binom{n+m}{k}$. Hence $X+Y\sim \mathrm{Bin}(n+m,p)$: a sum of independent binomials with the same success probability is again binomial. This is consistent with viewing $\mathrm{Bin}(n,p)$ as the number of successes in n independent $\mathrm{Ber}(p)$ trials (Appendix A.1).

We now turn to the behaviour of averages of many independent random variables. For i.i.d. $X_1, X_2, \dots$ with $Y_n := X_1+\cdots+X_n$, linearity of the expectation and Proposition 3.20 give

$$\mathbf{E}[Y_n] = n\,\mathbf{E}[X_1], \qquad \mathbf{Var}[Y_n] = n\,\mathbf{Var}[X_1]. \tag{3.1}$$

Thus the typical fluctuation of Y_n (of the order of its standard deviation) is of order $\sqrt{n}$.

Theorem 3.22 (Weak law of large numbers). *Let $X_1, X_2, \dots$ be real-valued, independent, identically distributed random variables with finite expectation $\mathbf{E}[X_1]=\mu$ and finite variance. Then for all $\varepsilon>0$,*

$$\lim_{n\to\infty}\mathbf{P}\Big(\Big|\frac{X_1+\cdots+X_n}{n}-\mu\Big|\ge\varepsilon\Big)=0.$$

Proof. By the Chebyshev inequality and (3.1),

$$\mathbf{P}\Big(\Big|\tfrac{1}{n}(X_1+\cdots+X_n)-\mu\Big|\ge\varepsilon\Big)\le\frac{\mathbf{Var}[\frac{1}{n}\sum_i X_i]}{\varepsilon^2}=\frac{\mathbf{Var}[X_1]}{n\varepsilon^2}\xrightarrow{n\to\infty}0.$$

□

3.5. **Conditional distribution and conditional expectation.** Recall the conditional distribution of Y given X from Section 2.3. Averaging the target function against this conditional distribution leads to the conditional expectation.

Definition 3.23 (Conditional expectation). *Let X, Y be random variables and $B\mapsto\mathbf{P}(Y\in B\,|\,X)$ the conditional distribution of Y given X. For a function h, the expectation of $h(Y)$ with respect to this conditional distribution is called the* conditional expectation of $h(Y)$ given X. *In the discrete case*

$$\mathbf{E}[h(Y)\,|\,X]=\sum_y h(y)\,\mathbf{P}(Y=y\,|\,X),$$

in the continuous case

$$\mathbf{E}[h(Y)\,|\,X]=\frac{\int h(y)f(X,y)\,dy}{\int f(X,y)\,dy},$$

provided the expectations exist.

Proposition 3.24 (Tower property). *Let X, Y be random variables and h a function. Then $\mathbf{E}\big[\mathbf{E}[h(Y)\,|\,X]\big]=\mathbf{E}[h(Y)]$.*

Proof. In the discrete case,

$$\mathbf{E}\big[\mathbf{E}[h(Y)\,|\,X]\big]=\sum_x\mathbf{E}[h(Y)\,|\,X=x]\cdot\mathbf{P}(X=x)=\sum_{x,y}h(y)\,\mathbf{P}(Y=y,X=x)=\mathbf{E}[h(Y)].$$

□

3.6. **The number of comparisons in** QUICKSORT**.** As an example of a genuinely complicated discrete random variable, we now gather in one place the analysis of the number of comparisons made by QUICKSORT on random input, drawing on the expectation, variance, and concentration tools of the previous section.

The model and a useful representation. The input is a vector of distinct numbers $x_1,\dots,x_n$, to be sorted in increasing order. Recall the QUICKSORT algorithm (Algorithm 1): choose $a := x_1$ as pivot, compare $x_2,\dots,x_n$ with a, and rearrange the vector as $((x_{j_1},\dots,x_{j_k}),\,a,\,(x_{i_1},\dots,x_{i_{n-k-1}}))$ with $x_{j_\bullet}<a<x_{i_\bullet}$, then recurse on the left and right parts. For example, sorting $(17,4,32,19,8,65,44)$ runs as follows: the pivot 17 is compared with the other six entries (6 comparisons) and splits the vector into $(4,8)$ and $(32,19,65,44)$; sorting $(4,8)$ compares its pivot 4 with 8 (1 comparison); sorting $(32,19,65,44)$ compares its pivot 32 with $19,65,44$ (3 comparisons) and leaves (19) and $(65,44)$; and sorting $(65,44)$ compares 65 with 44 (1 comparison). Singletons need no comparison, so in total $6+1+3+1=11$ comparisons are made.

Algorithm 1 QUICKSORT

INPUT: Unsorted vector $(x_1, \dots, x_n)$
OUTPUT: Sorted vector
1: **if** $n == 1$ **then** ans $= (x_1)$
2: **else**
3: $a \leftarrow x_1$
4: $(x_1, \dots, x_n) \leftarrow$ vector with all elements smaller (larger) than a to the left (right) of a
5: $q \leftarrow$ position of a in $(x_1, \dots, x_n)$
6: ans $=$ cat(QUICKSORT$(x_1, \dots, x_{q-1})$, a, QUICKSORT$(x_{q+1}, \dots, x_n)$)
7: **end if**
8: **return** ans

We study QUICKSORT with randomised input: for given numbers $z_1 < \cdots < z_n$ the input is a uniformly random reordering $z \circ \Sigma = (z_{\Sigma(1)}, \dots, z_{\Sigma(n)})$, i.e. $\mathbf{P}(\Sigma = \sigma) = 1/n!$ for $\sigma \in \mathcal{S}_n$. Then

$$h : \mathcal{S}_n \to \mathbb{N}, \qquad \sigma \mapsto \#\{\text{comparisons of QUICKSORT on input } z \circ \sigma\}$$

is a discrete random variable, the image of the uniform Σ under h. Without loss of generality we take $z_i = i$, so the input is a random permutation of $\{1, \dots, n\}$. For $i < j$ set $A_{ij} := \{\sigma : i \text{ and } j \text{ are compared}\}$ and $\gamma_{ij}(\sigma) := 1_{A_{ij}}(\sigma)$. Since i and j are compared at most once (exactly when one of them is the pivot),

$$(*) \qquad h(\sigma) = \sum_{i<j} \gamma_{ij}(\sigma).$$

As an aside, QUICKSORT needs the maximal $6 + 5 + \cdots + 1 = 21$ comparisons for $n = 7$ exactly when every pivot is the smallest or largest element of its subarray, that is on $2^{n-1} = 2^6$ of the inputs, e.g. $\mathbf{P}(h(\Sigma) = 21) = 2^6/7! = 64/7!$.

Expected number of comparisons. One sees that i and j are compared exactly when, in the sub-vector of σ containing the numbers $i, \dots, j$, either i or j comes first; hence

$$(\square) \qquad \gamma_{ij}(\sigma) = 1(\text{in the sub-vector of } \sigma \text{ containing } i, \dots, j,\ i \text{ or } j \text{ comes first}).$$

Since the probability that i or j comes first in a permutation of $i, \dots, j$ is $\frac{2}{j-i+1}$, we get $\mathbf{P}(\Sigma \in A_{ij}) = \frac{2}{j-i+1}$, and, by Example 3.9 and Proposition 3.11,

$$\mathbf{E}[h(\Sigma)] = \sum_{i<j} \mathbf{E}[\gamma_{ij}(\Sigma)] = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{2}{j-i+1} = \Theta\Big(\int_1^n 2\log(x)\, dx \Big) = \Theta\big(2n \log n\big).$$

Thus QUICKSORT has an average running time (which is presumably proportional to the number of comparisons) of $\Theta(2n \log n)$.[3]

A second derivation, by conditioning on the pivot. The same expectation can be obtained by conditioning on the pivot and using the tower property (Proposition 3.24). Write $c_n := \mathbf{E}[h(\Sigma)]$ for a uniformly random $\Sigma \in \mathcal{S}_n$. The pivot $z_{\Sigma(1)}$ has rank $\Sigma(1)$, which is uniform on $\{1, \dots, n\}$. Given $\Sigma(1) = r$, QUICKSORT first compares the pivot with the other $n - 1$ elements and then sorts the $r - 1$ elements below it and the $n - r$ above it; conditioned on the rank, each part is itself in uniformly random order, so

$$\mathbf{E}[h(\Sigma) \,|\, \Sigma(1) = r] = (n-1) + c_{r-1} + c_{n-r}.$$

By the tower property, and since $\sum_{r=1}^{n} c_{r-1} = \sum_{r=1}^{n} c_{n-r} = \sum_{i=0}^{n-1} c_i$,

$$c_n = \frac{1}{n} \sum_{r=1}^{n} \big((n-1) + c_{r-1} + c_{n-r} \big) = (n-1) + \frac{2}{n} \sum_{i=0}^{n-1} c_i, \qquad c_0 = c_1 = 0.$$

[3] We write $g(n) = \Theta(f(n))$ to mean that g and f are of the same order of magnitude: there are constants $0 < c_1 \le c_2$ and an n_0 with $c_1 f(n) \le g(n) \le c_2 f(n)$ for all $n \ge n_0$.

Solving this recurrence (multiply by n, subtract the same identity for $n-1$, and telescope) gives $c_n = 2(n+1)H_n - 4n$ with $H_n = \sum_{k=1}^{n} \frac{1}{k}$. This is precisely the exact value of the double sum above: grouping the pairs by their gap $d = j - i$ (there are $n - d$ pairs with $j - i = d$),

$$\sum_{i<j} \frac{2}{j-i+1} = \sum_{d=1}^{n-1} (n-d)\frac{2}{d+1} = 2(n+1)H_n - 4n.$$

Since $H_n = \log n + O(1)$, both derivations give $\mathbf{E}[h(\Sigma)] = \Theta(2n \log n)$.
Variance. We next bound $\mathbf{Var}[h(\Sigma)] \le 3n(n-1)$. By Proposition 3.14 and $1_A^2 = 1_A$,

$$\mathbf{Var}[\gamma_{ij}(\Sigma)] = \frac{2}{j-i+1} - \Big(\frac{2}{j-i+1}\Big)^2 \le \frac{2}{j-i+1}.$$

For the covariance terms set $T_{ij}(\sigma) :=$ the first entry in the sub-vector $i, \dots, j$ of σ, so that, by $(\square)$, $\gamma_{ij} = 1$ iff $T_{ij} \in \{i, j\}$. For $i < j \ne k < l$ (disjoint index blocks) T_{ij} and T_{kl} are independent, so the covariance vanishes. The overlapping cases $j = k$, $i = k$, $j = l$ each contribute at most $\frac{1}{l-i+1}$ beyond the product of the means. With Proposition 3.20 (the factor 6 combining a factor 2 and the three cases) and the representation $(*)$,

$$\mathbf{Var}[h(\Sigma)] \le \sum_{i<j} \frac{2}{j-i+1} + \sum_{i<j<l} \frac{6}{l-i+1} \le \sum_{i<j} 6 = 6\binom{n}{2} = 3n(n-1).$$

Concentration. A law-of-large-numbers effect can also occur for dependent random variables. The $\gamma_{ij}(\Sigma)$ are neither identically distributed (their mean depends on $j - i$) nor independent (some covariances are non-zero, as just seen). Nevertheless, by the Chebyshev inequality together with the mean and variance found above,

$$\mathbf{P}\Big(\Big|\frac{h(\Sigma)}{\mathbf{E}[h(\Sigma)]} - 1\Big| > \varepsilon\Big) \le \frac{1}{\varepsilon^2} \frac{\mathbf{Var}[h(\Sigma)]}{\mathbf{E}[h(\Sigma)]^2} \le \frac{1}{\varepsilon^2} \frac{3n(n-1)}{\Theta(4n^2(\log n)^2)} \xrightarrow{n\to\infty} 0.$$

So the number of comparisons concentrates around its mean $\Theta(2n \log n)$.

4. The need of continuous random variables

So far we have mostly worked with discrete random variables. Many quantities, however, are naturally continuous: waiting times, physical measurements, or the numbers in $[0, 1)$ that a random-number generator is supposed to produce. Such quantities cannot be captured by a probability mass function, and we need densities.

4.1. **Densities.** A first hint of this need comes from a limit of discrete distributions. Consider the geometric distribution geo(p) (Appendix A.5), the waiting time for the first success in a p-coin toss, with $\mathbf{P}(X = k) = (1-p)^{k-1}p$ and $\mathbf{P}(X > k) = (1-p)^k$. Waiting for the r-th success instead of the first is the natural generalisation and leads to the *negative binomial distribution* NB(r, p) (Appendix A.6).

Lemma 4.1 (Exponential and gamma approximation). *Let $p_n \to 0$ as $n \to \infty$, and fix $r \in \mathbb{N}$.*

(1) *If $X_n \sim$ geo(p_n), then for every $x \ge 0$,*

$$\lim_{n\to\infty} \mathbf{P}(p_n X_n > x) = e^{-x};$$

that is, the rescaled waiting time $p_n X_n$ converges in distribution to the exponential distribution exp(1) *(Appendix A.10).*

(2) *If $X_n \sim$ NB(r, p_n), then $p_n X_n$ converges in distribution to the gamma distribution $\Gamma(r, 1)$ (Appendix A.12), the law of a sum of r independent* exp(1) *waiting times.*

Proof. 1. $\mathbf{P}(p_n X_n > x) = \mathbf{P}(X_n > x/p_n) = (1-p_n)^{\lfloor x/p_n \rfloor} \xrightarrow{n\to\infty} e^{-x}$.
2. Write $X_n = Y_1^{(n)} + \cdots + Y_r^{(n)}$ with independent $Y_i^{(n)} \sim$ geo(p_n) (Appendix A.6). By part 1 each $p_n Y_i^{(n)}$ converges to an exp(1) variable, and, the $Y_i^{(n)}$ being independent, the vector $(p_n Y_1^{(n)}, \dots, p_n Y_r^{(n)})$ converges to $(E_1, \dots, E_r)$ with independent $E_i \sim$ exp(1). Hence $p_n X_n = \sum_{i=1}^{r} p_n Y_i^{(n)}$ converges to $E_1 + \cdots + E_r \sim \Gamma(r, 1)$. □

The rescaled waiting time $p_n X_n$ therefore has, in the limit, a distribution that is no longer discrete: its "mass" is spread out over $[0,\infty)$ and described by a density. This motivates the following notion, already announced in Definition 1.7.2.

Definition 4.2 (Density). *A function $f:\mathbb{R}\to[0,\infty)$ with*

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

is called a density. *A real-valued random variable X is* continuous *with density f if*

$$\mathbf{P}(X\in A) = \int_A f(x)\,dx \qquad \text{for every interval } A;$$

we write $\mathbf{P}(X\in dx) = f(x)\,dx$. Its distribution function is then

$$F_X(x) = \int_{-\infty}^{x} f(z)\,dz, \qquad \text{so that } F_X' = f.$$

Remark 4.3 (Expectation and variance in the continuous case). All the computational rules for expectation and variance from Section 3.3 carry over unchanged to continuous random variables; one only replaces sums against the probability mass function by integrals against the density, and no new proof is needed. The formula

$$\mathbf{E}[h(X)] = \int_{-\infty}^{\infty} h(x)\,f(x)\,dx$$

is already part of the definition of the expectation. Taking $h(x)=x$ and $h(x)=(x-\mu)^2$ (with $\mu := \mathbf{E}[X]$) gives

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x\,f(x)\,dx, \qquad \mathbf{Var}[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \mathbf{E}[X^2]-\mu^2$$

(Proposition 3.14). Linearity of the expectation and $\mathbf{Var}[aX+b] = a^2\,\mathbf{Var}[X]$ (Proposition 3.11) continue to hold, as do the tail formulae

$$\mathbf{E}[X] = \int_0^{\infty} \mathbf{P}(X>x)\,dx, \qquad \mathbf{E}[X^2] = 2\int_0^{\infty} x\,\mathbf{P}(X>x)\,dx$$

(Lemma 3.10) and the Markov and Chebyshev inequalities — their proofs never used discreteness.

The concrete continuous distributions used in these notes — the uniform $U([a,b])$, the exponential $\exp(\lambda)$, the normal $N(\mu,\sigma^2)$, and the gamma $\Gamma(\alpha,\lambda)$ — are collected, with their densities, distribution functions, expectations, and variances, in the catalogue of Appendix A. Among them, the normal $N(\mu,\sigma^2)$ (Appendix A.11) is the limiting shape in the central limit theorem, which we treat next (Theorem 4.4).

4.2. **The central limit theorem.** The central limit theorem refines the weak law of large numbers (Theorem 3.22): whereas the latter says that the sample mean concentrates at μ, the former describes the size and shape of the fluctuations around μ, on the scale $1/\sqrt{n}$. Remarkably, the limiting shape is always the normal distribution, whatever the distribution of the summands.

Theorem 4.4 (Central limit theorem). *Let $X_1, X_2, \dots$ be independent, identically distributed, real-valued random variables with finite expectation μ and finite variance $\sigma^2>0$. Then, for all $-\infty\le c<d\le\infty$,*

$$\lim_{n\to\infty} \mathbf{P}\Big(c \le \frac{X_1+\cdots+X_n-n\mu}{\sqrt{n\sigma^2}} \le d\Big) = \mathbf{P}(c\le Z\le d), \qquad Z\sim N(0,1).$$

Remark 4.5.

(1) The standardised sum $\frac{X_1+\cdots+X_n-n\mu}{\sqrt{n\sigma^2}}$ has expectation 0 and variance 1, just like Z; the theorem asserts that its *whole* distribution converges to $N(0,1)$. (For a single random variable X, the map $X\mapsto \frac{X-\mathbf{E}[X]}{\sqrt{\mathbf{Var}[X]}}$ is called its *standardisation*.)

(2) If already $X_1, X_2, \cdots \sim N(\mu,\sigma^2)$, then $X_1+\cdots+X_n \sim N(n\mu, n\sigma^2)$, so the standardised sum is exactly $N(0,1)$-distributed and the theorem holds even without the limit.

(3) For $X_1, X_2, \dots \sim \mathrm{Ber}(p)$, so that $S_n := X_1 + \dots + X_n \sim \mathrm{Bin}(n,p)$, the statement

$$\mathbf{P}\Big(c \le \frac{S_n - np}{\sqrt{np(1-p)}} \le d\Big) \xrightarrow{n\to\infty} \mathbf{P}(c \le Z \le d)$$

is known as the *theorem of de Moivre–Laplace*.

Before proving the theorem we isolate its analytic core: for smooth test functions the normalised sum may be compared, term by term, with a sum of normal random variables. This replacement trick goes back to Lindeberg.

Lemma 4.6. *Let $X_1, X_2, \dots$ be as in Theorem 4.4 with $\mu = 0$ and $\sigma^2 = 1$, let $Z \sim N(0,1)$, and let $h : \mathbb{R} \to \mathbb{R}$ be three times continuously differentiable with bounded first three derivatives. Then*

$$\mathbf{E}\Big[h\Big(\frac{X_1 + \dots + X_n}{\sqrt{n}}\Big)\Big] \xrightarrow{n\to\infty} \mathbf{E}[h(Z)].$$

Proof. Let $Z_1, Z_2, \dots \sim N(0,1)$ be independent and independent of $X_1, X_2, \dots$, and set $c_2 := \sup_x |h''(x)|$ and $c_3 := \sup_x |h'''(x)|$. For $i = 1, \dots, n$ put

$$U_i := \frac{X_1 + \dots + X_{i-1} + Z_{i+1} + \dots + Z_n}{\sqrt{n}}.$$

Replacing the Z_i by the X_i one at a time and telescoping,

$$h\Big(\frac{X_1 + \dots + X_n}{\sqrt{n}}\Big) - h\Big(\frac{Z_1 + \dots + Z_n}{\sqrt{n}}\Big) = \sum_{i=1}^{n} \Big[h\Big(U_i + \tfrac{X_i}{\sqrt{n}}\Big) - h\Big(U_i + \tfrac{Z_i}{\sqrt{n}}\Big)\Big], \tag{4.1}$$

since $U_i + X_i/\sqrt{n}$ and $U_{i+1} + Z_{i+1}/\sqrt{n}$ are the same random variable. By Taylor's formula, for each i

$$h\Big(U_i + \frac{X_i}{\sqrt{n}}\Big) - h\Big(U_i + \frac{Z_i}{\sqrt{n}}\Big) = h'(U_i)\frac{X_i - Z_i}{\sqrt{n}} + h''(U_i)\frac{X_i^2 - Z_i^2}{2n} + R_{in},$$

where, with intermediate points V_i, W_i satisfying $|V_i - U_i| \le |X_i|/\sqrt{n}$ and $|W_i - U_i| \le |Z_i|/\sqrt{n}$,

$$R_{in} = \big(h''(V_i) - h''(U_i)\big)\frac{X_i^2}{2n} + \big(h''(U_i) - h''(W_i)\big)\frac{Z_i^2}{2n}.$$

Bounding $|h''(V_i) - h''(U_i)|$ either by $c_3|V_i - U_i| \le c_3|X_i|/\sqrt{n}$ or by $2c_2$ (and likewise for W_i), we get, for arbitrary $k > 0$,

$$|R_{in}| \le c_3\frac{k^3}{n^{3/2}} + c_2\frac{X_i^2}{n}1_{|X_i|>k} + c_3\frac{|Z_i|^3}{n^{3/2}}.$$

Now U_i is independent of (X_i, Z_i), and X_i, Z_i share the first two moments, $\mathbf{E}[X_i] = \mathbf{E}[Z_i] = 0$ and $\mathbf{E}[X_i^2] = \mathbf{E}[Z_i^2] = 1$. Hence the first two terms above have expectation 0,

$$\mathbf{E}\Big[h'(U_i)\frac{X_i - Z_i}{\sqrt{n}}\Big] = 0, \qquad \mathbf{E}\Big[h''(U_i)\frac{X_i^2 - Z_i^2}{2n}\Big] = 0,$$

and therefore

$$\Big|\mathbf{E}\Big[h\Big(U_i + \tfrac{X_i}{\sqrt{n}}\Big) - h\Big(U_i + \tfrac{Z_i}{\sqrt{n}}\Big)\Big]\Big| = |\mathbf{E}[R_{in}]| \le c_3\frac{k^3 + \mathbf{E}[|Z_1|^3]}{n^{3/2}} + c_2\frac{\mathbf{E}[X_1^2 1_{|X_1|>k}]}{n}.$$

Summing over i in (4.1) and using that $(Z_1 + \dots + Z_n)/\sqrt{n} \sim N(0,1)$,

$$\Big|\mathbf{E}\Big[h\Big(\frac{X_1 + \dots + X_n}{\sqrt{n}}\Big)\Big] - \mathbf{E}[h(Z)]\Big| \le c_3\frac{k^3 + \mathbf{E}[|Z_1|^3]}{n^{1/2}} + c_2\,\mathbf{E}[X_1^2 1_{|X_1|>k}].$$

Given $\varepsilon > 0$, first choose k so large that $c_2\,\mathbf{E}[X_1^2 1_{|X_1|>k}] < \varepsilon$ (possible since $\mathbf{E}[X_1^2] = 1 < \infty$), then let $n \to \infty$, so that the first term vanishes. As $\varepsilon > 0$ was arbitrary, the claim follows. □

Proof of Theorem 4.4. Passing from X_i to $(X_i - \mu)/\sigma$, we may assume $\mu = 0$ and $\sigma^2 = 1$; write $Y_n^* := (X_1 + \dots + X_n)/\sqrt{n}$. Fix $\varepsilon > 0$ and choose functions h_1, h_2, three times continuously differentiable with bounded derivatives, such that

$$1_{[c+\varepsilon,\, d-\varepsilon]} \le h_1 \le 1_{[c,d]} \le h_2 \le 1_{[c-\varepsilon,\, d+\varepsilon]}.$$

By Lemma 4.6, $\mathbf{E}[h_j(Y_n^*)] \to \mathbf{E}[h_j(Z)]$ for $j = 1, 2$, whence

$$\begin{aligned}\mathbf{P}(c+\varepsilon \le Z \le d-\varepsilon) &\le \mathbf{E}[h_1(Z)] = \lim_{n\to\infty} \mathbf{E}[h_1(Y_n^*)] \le \liminf_{n\to\infty} \mathbf{P}(c \le Y_n^* \le d) \\ &\le \limsup_{n\to\infty} \mathbf{P}(c \le Y_n^* \le d) \le \lim_{n\to\infty} \mathbf{E}[h_2(Y_n^*)] = \mathbf{E}[h_2(Z)] \\ &\le \mathbf{P}(c-\varepsilon \le Z \le d+\varepsilon).\end{aligned}$$

Letting $\varepsilon \to 0$ and using that $\mathbf{P}(c \pm \varepsilon \le Z \le d \pm \varepsilon) \to \mathbf{P}(c \le Z \le d)$ (the standard normal has no atoms), both the lower and the upper limit equal $\mathbf{P}(c \le Z \le d)$, which is the assertion. □

Appendix A. A catalogue of distributions

For each distribution we describe a random experiment leading to it, give its distribution (probability mass function or density, and the distribution function where it is available in closed form), and compute its expectation and variance. Throughout, $q := 1 - p$, and 1_A denotes an indicator. We repeatedly use the identities $\mathbf{E}[X] = \sum_{x \ge 0} \mathbf{P}(X > x)$ and $\mathbf{Var}[X] = \mathbf{E}[X(X-1)] + \mathbf{E}[X] - \mathbf{E}[X]^2$ (Lemma 3.10, Proposition 3.14), and, for sums of uncorrelated variables, $\mathbf{Var}[\sum_i X_i] = \sum_i \mathbf{Var}[X_i]$ (Proposition 3.20).

A.1. **Bernoulli distribution** $\mathrm{Ber}(p)$**. Experiment.** A single trial with success probability $p \in [0,1]$ (one coin toss, one bit): $X = 1$ on success, $X = 0$ on failure.
Distribution.

$$\mathbf{P}(X = 1) = p, \qquad \mathbf{P}(X = 0) = q.$$

Expectation and variance. Since $X^2 = X$,

$$\mathbf{E}[X] = 0 \cdot q + 1 \cdot p = p, \qquad \mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = p - p^2 = pq.$$

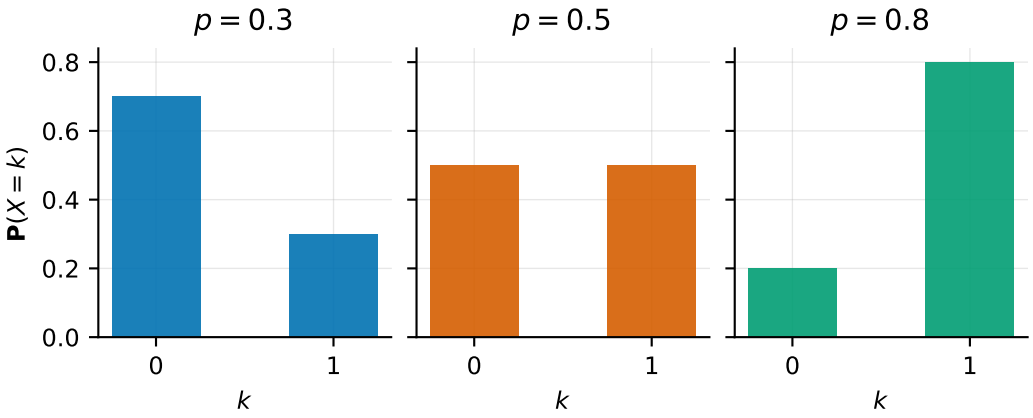

Figure A.1. Probability mass function of $\mathrm{Ber}(p)$ for three values of p.

A.2. **Binomial distribution** $\mathrm{Bin}(n,p)$**. Experiment.** Count the number X of successes in n independent $\mathrm{Ber}(p)$ trials (e.g. the number of heads in an n-fold p-coin toss).
Distribution.

$$\mathbf{P}(X = k) = \binom{n}{k} p^k q^{n-k}, \qquad k = 0, \dots, n.$$

Expectation and variance. Directly from the mass function, using $k\binom{n}{k} = n\binom{n-1}{k-1}$,

$$\mathbf{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} = np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j q^{n-1-j} = np,$$

the remaining sum being 1 by the binomial theorem. Likewise, from $k(k-1)\binom{n}{k} = n(n-1)\binom{n-2}{k-2}$,

$$\mathbf{E}[X(X-1)] = \sum_{k=0}^{n} k(k-1) \binom{n}{k} p^k q^{n-k} = n(n-1)p^2 \sum_{j=0}^{n-2} \binom{n-2}{j} p^j q^{n-2-j} = n(n-1)p^2,$$

so that, by $\mathbf{Var}[X] = \mathbf{E}[X(X-1)] + \mathbf{E}[X] - \mathbf{E}[X]^2$,

$$\mathbf{Var}[X] = n(n-1)p^2 + np - (np)^2 = npq.$$

(Equivalently, writing $X = X_1 + \cdots + X_n$ with independent $X_i \sim \mathrm{Ber}(p)$ gives $\mathbf{E}[X] = np$ and, by uncorrelatedness (Proposition 3.19), $\mathbf{Var}[X] = npq$ at once.)

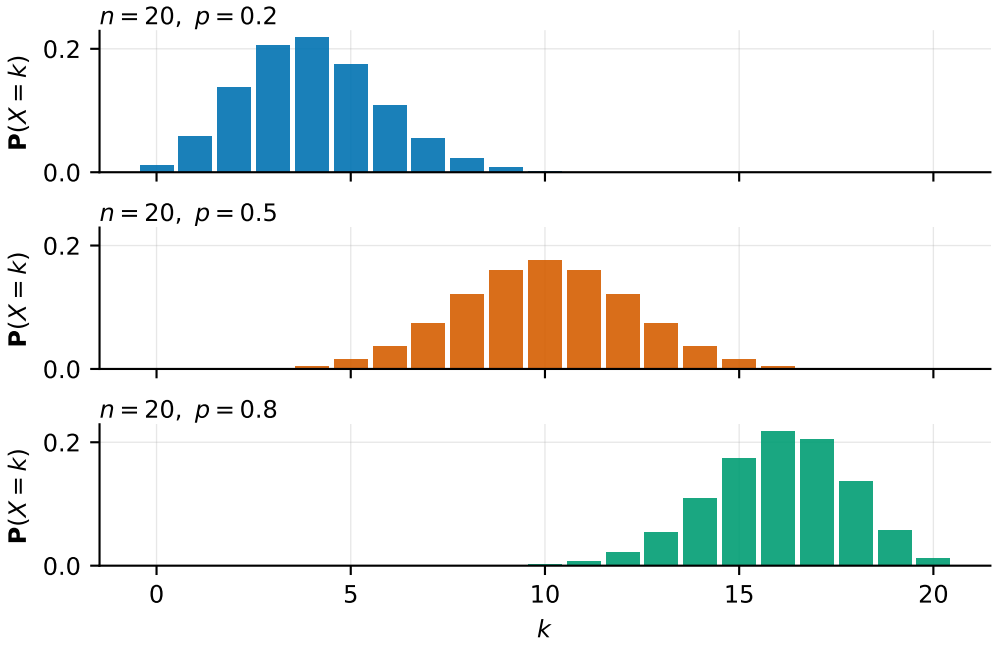

Figure A.2. Probability mass function of $\mathrm{Bin}(n,p)$ for $n = 20$ and three values of p.

A.3. **Multinomial distribution** $\mathrm{Mult}(n; p_1, \dots, p_\ell)$**. Experiment.** Perform n independent trials, each of which falls into one of ℓ categories, category i with probability p_i (where $p_1 + \cdots + p_\ell = 1$); let X_i be the number of trials in category i. For instance, distribute n keys over ℓ hash bins and count how many land in each bin. This is the multivariate generalisation of the binomial (which is the case $\ell = 2$).

Distribution. For $(k_1, \dots, k_\ell)$ with $k_1 + \cdots + k_\ell = n$,

$$\mathbf{P}\big(X = (k_1, \dots, k_\ell)\big) = \binom{n}{k_1 \ \cdots \ k_\ell} p_1^{k_1} \cdots p_\ell^{k_\ell}, \qquad \binom{n}{k_1 \ \cdots \ k_\ell} := \frac{n!}{k_1! \cdots k_\ell!}.$$

Expectation and variance. Each coordinate X_i counts the successes of category i against all others, so $X_i \sim \mathrm{Bin}(n, p_i)$ and

$$\mathbf{E}[X_i] = np_i, \qquad \mathbf{Var}[X_i] = np_i(1-p_i).$$

For $i \neq j$ we use the indicator decomposition $X_i = \sum_{m=1}^n A_m^{(i)}$ with $A_m^{(i)} := 1_{\{\text{trial } m \text{ in category } i\}}$, so that

$$\mathbf{Cov}[X_i, X_j] = \sum_{m=1}^n \sum_{m'=1}^n \mathbf{Cov}\big[A_m^{(i)}, A_{m'}^{(j)}\big].$$

For $m \neq m'$ the trials m and m' are independent, so the covariance vanishes. For $m = m'$ a single trial cannot lie in two categories at once, hence $A_m^{(i)} A_m^{(j)} = 0$ and

$$\mathbf{Cov}\big[A_m^{(i)}, A_m^{(j)}\big] = \mathbf{E}\big[A_m^{(i)} A_m^{(j)}\big] - p_i p_j = -p_i p_j.$$

Summing the n diagonal terms gives

$$\mathbf{Cov}[X_i, X_j] = -np_i p_j.$$

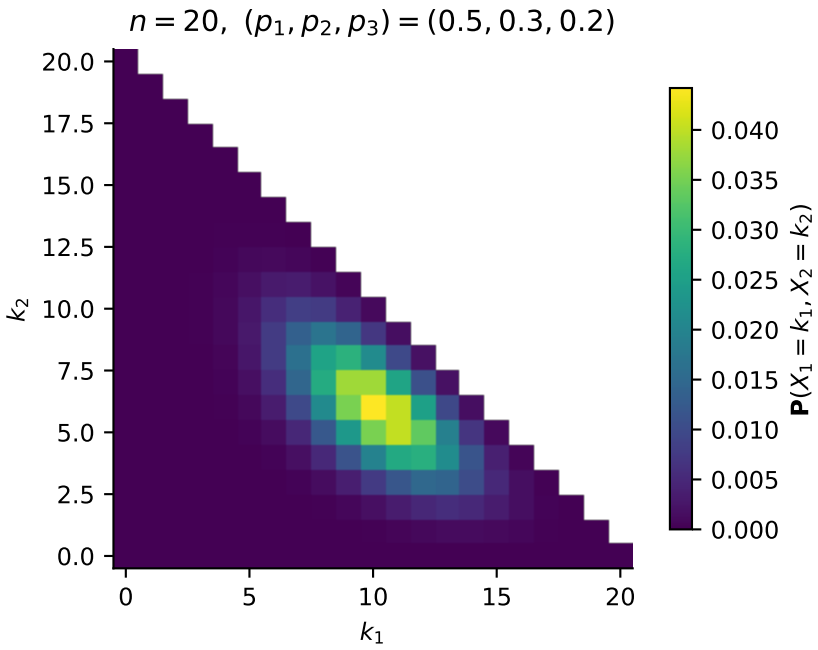

Figure A.3. Joint probability mass function of (X_1, X_2) for $\mathrm{Mult}(n; p_1, p_2, p_3)$ with $n = 20$ and $(p_1, p_2, p_3) = (0.5, 0.3, 0.2)$; here $X_3 = n - X_1 - X_2$ is determined. The support is the triangle $k_1 + k_2 \le n$.

A.4. **Hypergeometric distribution** $\mathrm{Hyp}(n, g, w)$**. Experiment.** An urn contains g balls, of which w are white. Draw n balls *without* replacement and let X be the number of white balls drawn (e.g. the winning numbers in a lottery).

Distribution.

$$\mathbf{P}(X = k) = \frac{\binom{w}{k}\binom{g-w}{n-k}}{\binom{g}{n}}, \qquad k = 0, \dots, n.$$

Expectation and variance. With $p := w/g$, let $Z_i := 1_{\{i\text{-th drawn ball is white}\}}$, so $X = \sum_{i=1}^{n} Z_i$. Imagine drawing *all* g balls in a uniformly random order, the first n of which form our sample. By symmetry the ball in a given position i is uniform among all g balls, and the ordered pair of balls in two positions $i \neq j$ is uniform among all $g(g-1)$ ordered pairs of distinct balls. Hence

$$\mathbf{P}(Z_i = 1) = \frac{w}{g} = p, \qquad \mathbf{P}(Z_i = Z_j = 1) = \frac{w(w-1)}{g(g-1)} \quad (i \neq j),$$

and so $\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{P}(Z_i = 1) = np$. For $i \neq j$,

$$\mathbf{Cov}[Z_i, Z_j] = \mathbf{P}(Z_i = Z_j = 1) - p^2 = p\Big(\frac{w-1}{g-1} - \frac{w}{g}\Big) = -\frac{pq}{g-1},$$

so by Proposition 3.20

$$\mathbf{Var}[X] = npq - n(n-1)\frac{pq}{g-1} = npq\Big(1 - \frac{n-1}{g-1}\Big).$$

(For $n \ll g$ this is close to the binomial value npq obtained when drawing *with* replacement.)

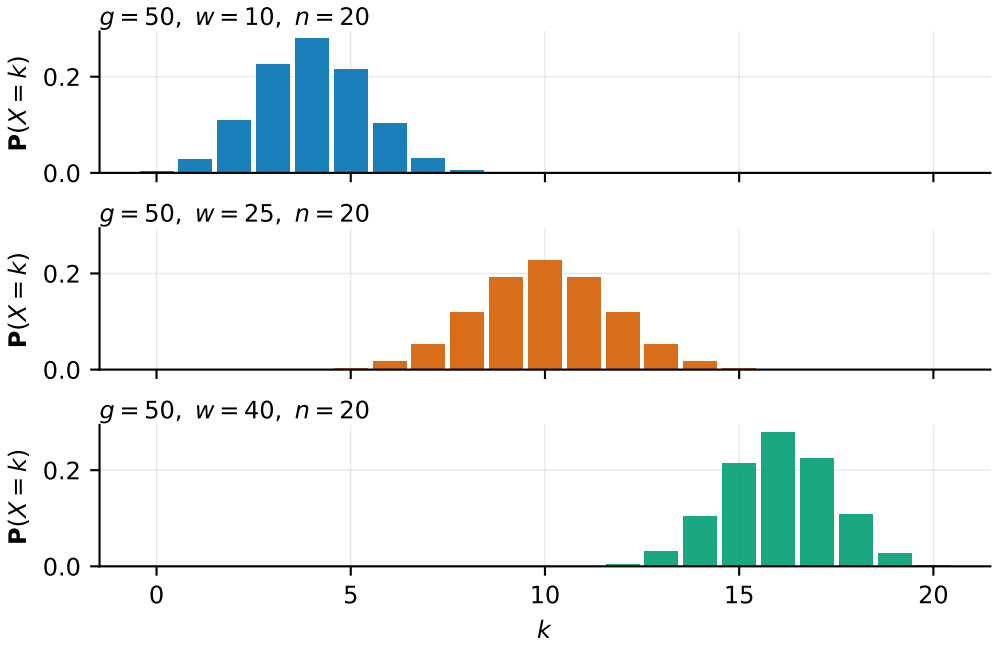

Figure A.4. Probability mass function of $\mathrm{Hyp}(n, g, w)$ for $g = 50$ balls, $n = 20$ draws, and three values of the number w of white balls.

A.5. **Geometric distribution** $\mathrm{geo}(p)$**. Experiment.** Repeat independent $\mathrm{Ber}(p)$ trials until the first success; let X be the number of trials needed.
Distribution.

$$\mathbf{P}(X = k) = q^{k-1}p, \quad k = 1, 2, \dots, \qquad \text{with tail } \mathbf{P}(X > k) = q^k.$$

Expectation and variance. Using the tail formulas,

$$\mathbf{E}[X] = \sum_{x=0}^{\infty} \mathbf{P}(X > x) = \sum_{x=0}^{\infty} q^x = \frac{1}{p},$$

$$\mathbf{E}[X(X-1)] = 2\sum_{x=0}^{\infty} x\, q^x = 2\frac{q}{p}\,\mathbf{E}[X] = \frac{2q}{p^2},$$

hence

$$\mathbf{Var}[X] = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{q}{p^2}.$$

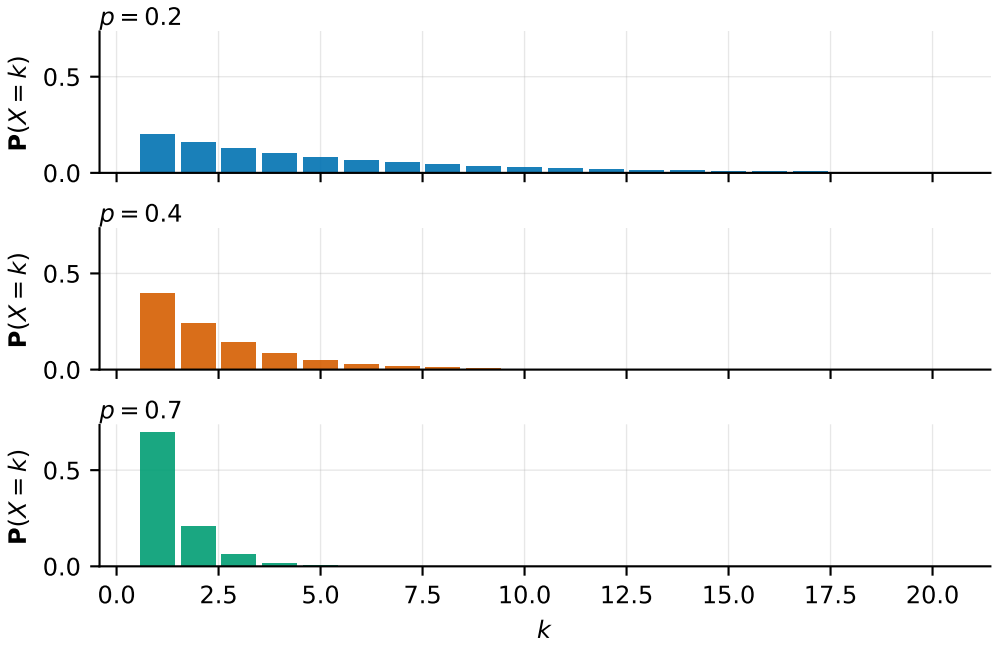

Figure A.5. Probability mass function of geo(p) (number of trials until the first success) for three values of p.

A.6. **Negative binomial distribution** NB(r, p)**. Experiment.** Repeat independent Ber(p) trials until the r-th success; let X be the number of trials needed.
Distribution. The last trial is a success, and among the first $k-1$ trials there are exactly $r-1$ successes, so

$$\mathbf{P}(X = k) = \binom{k-1}{r-1} p^r q^{k-r}, \qquad k = r, r+1, \dots$$

Expectation and variance. $X = Y_1 + \cdots + Y_r$, where Y_i is the number of trials between the $(i-1)$-th and the i-th success; the Y_i are independent and geo(p)-distributed. Hence

$$\mathbf{E}[X] = r\,\mathbf{E}[Y_1] = \frac{r}{p}, \qquad \mathbf{Var}[X] = r\,\mathbf{Var}[Y_1] = \frac{rq}{p^2}.$$

(The case $r = 1$ is the geometric distribution.)

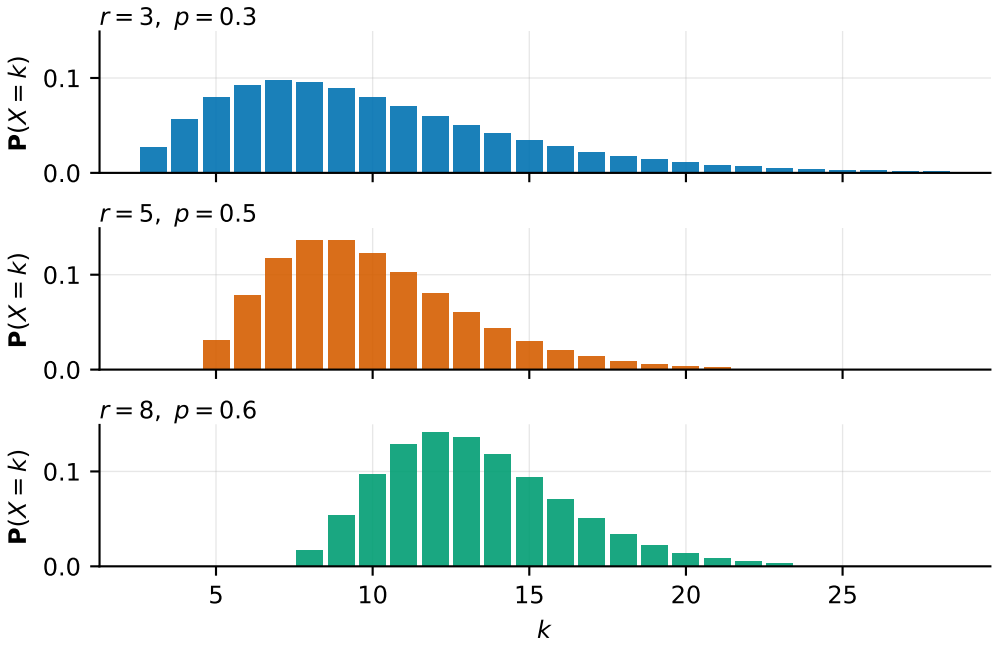

Figure A.6. Probability mass function of $\mathrm{NB}(r,p)$ (number of trials until the r-th success) for three parameter pairs.

A.7. **Poisson distribution** $\mathrm{Poi}(\lambda)$**. Experiment.** The number of "rare events": the limit of $\mathrm{Bin}(n,p_n)$ with $np_n \to \lambda$ (law of small numbers, Proposition 3.5); e.g. the number of jackpot winners in a lottery.

Distribution.

$$\mathbf{P}(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k=0,1,2,\dots$$

Expectation and variance. Since $\sum_{k\geq 0}\lambda^k/k! = e^{\lambda}$,

$$\mathbf{E}[X] = \sum_{k=0}^{\infty} k\, e^{-\lambda}\frac{\lambda^k}{k!} = \lambda, \qquad \mathbf{E}[X(X-1)] = \sum_{k=0}^{\infty} k(k-1)e^{-\lambda}\frac{\lambda^k}{k!} = \lambda^2,$$

hence

$$\mathbf{Var}[X] = \lambda^2 + \lambda - \lambda^2 = \lambda, \qquad \text{so} \qquad \mathbf{E}[X] = \mathbf{Var}[X] = \lambda.$$

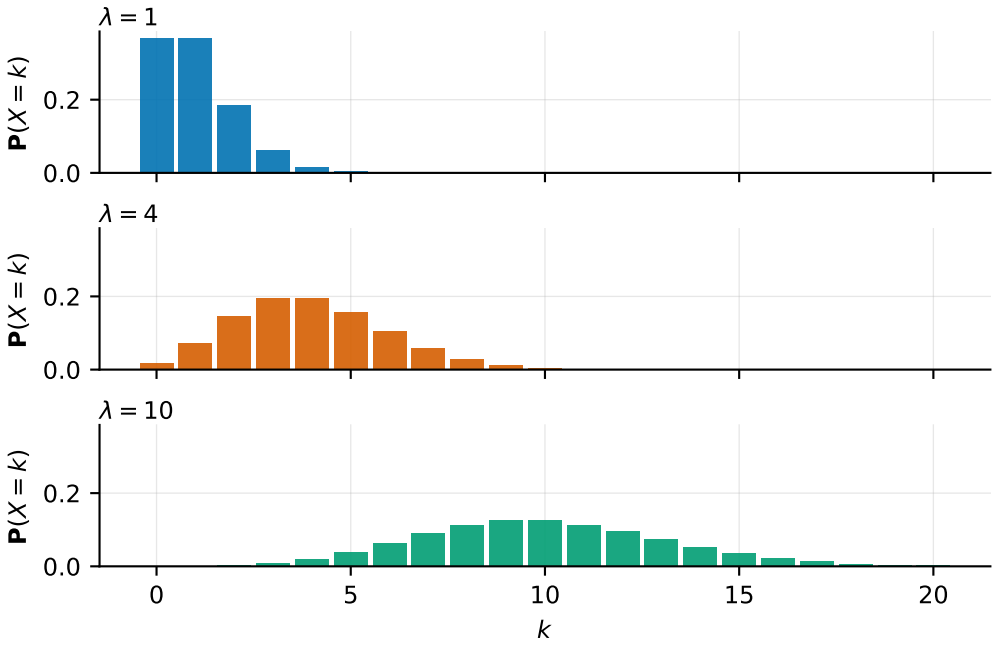

Figure A.7. Probability mass function of Poi(λ) for three values of λ.

A.8. **Discrete uniform distribution. Experiment.** Pick an element uniformly at random from a finite set E (a fair die, a random permutation, ...).
Distribution.

$$\mathbf{P}(X = x) = \frac{1}{|E|} \quad \text{for } x \in E, \qquad \mathbf{P}(X \in A) = \frac{|A|}{|E|}.$$

Expectation and variance. For $E = \{k, k+1, \dots, \ell\}$ with $m := \ell - k + 1$ equally likely values,

$$\mathbf{E}[X] = \frac{1}{m} \sum_{x=k}^{\ell} x = \frac{k+\ell}{2}.$$

A shift by $k-1$ leaves the variance unchanged, so X has the same variance as a uniform variable on $\{1, \dots, m\}$; with $\sum_{j=1}^{m} j^2 = \frac{m(m+1)(2m+1)}{6}$,

$$\mathbf{Var}[X] = \frac{1}{m} \sum_{j=1}^{m} j^2 - \Big(\frac{m+1}{2}\Big)^2 = \frac{m^2-1}{12} = \frac{(\ell-k+1)^2-1}{12}.$$

(In particular, for $E = \{1, \dots, n\}$ one has $\mathbf{E}[X] = \frac{n+1}{2}$ and $\mathbf{Var}[X] = \frac{n^2-1}{12}$.)

A.9. **Continuous uniform distribution $U([a,b])$. Experiment.** A point chosen uniformly at random in an interval $[a,b]$ (e.g. the output of an ideal random-number generator on $[0,1]$).
Distribution. The density and distribution function are

$$f(x) = \frac{1}{b-a} 1_{[a,b]}(x), \qquad F(x) = \frac{x-a}{b-a} \quad \text{for } a \le x < b$$

(and $F = 0$, 1 outside $[a,b]$).
Expectation and variance.

$$\mathbf{E}[X] = \frac{1}{b-a} \int_a^b x\, dx = \frac{a+b}{2}, \qquad \mathbf{Var}[X] = \frac{1}{b-a} \int_a^b x^2\, dx - \Big(\frac{a+b}{2}\Big)^2 = \frac{(b-a)^2}{12}.$$

A.10. **Exponential distribution** $\exp(\lambda)$**. Experiment.** A continuous, memoryless waiting time; the limit of a rescaled geometric waiting time (exponential approximation), e.g. the time until the next event.
Distribution. The density and distribution function are, for $x \ge 0$,

$$f(x) = \lambda e^{-\lambda x}, \qquad F(x) = 1 - e^{-\lambda x}.$$

Expectation and variance. Using the tail formulas (Lemma 3.10) with $\mathbf{P}(X > x) = e^{-\lambda x}$,

$$\mathbf{E}[X] = \int_0^\infty \mathbf{P}(X > x)\, dx = \int_0^\infty e^{-\lambda x}\, dx = \frac{1}{\lambda},$$
$$\mathbf{E}[X^2] = 2\int_0^\infty x\, \mathbf{P}(X > x)\, dx = 2\int_0^\infty x e^{-\lambda x}\, dx = \frac{2}{\lambda^2},$$

hence

$$\mathbf{Var}[X] = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

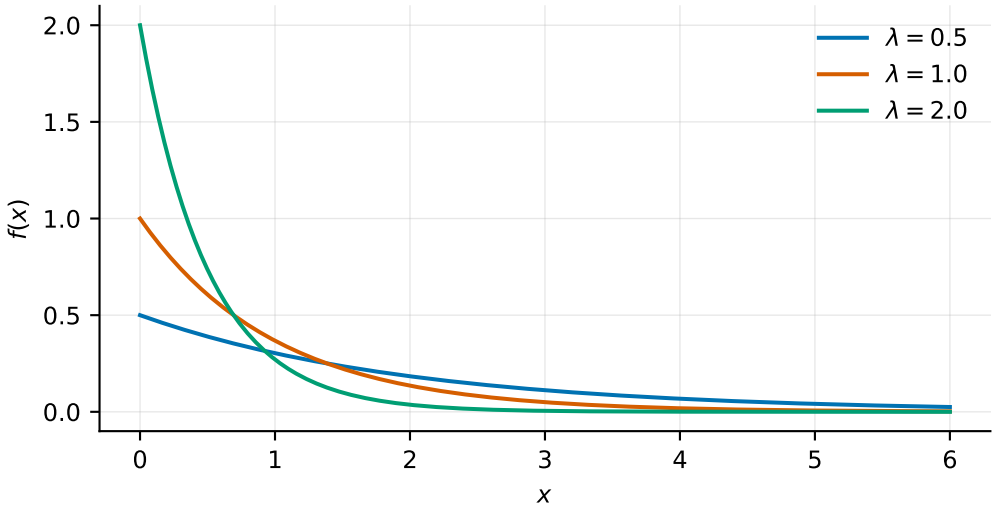

Figure A.8. Density of $\exp(\lambda)$ for three values of the rate λ.

A.11. **Normal distribution $N(\mu, \sigma^2)$. Experiment.** Measurement errors, or a sum of many small independent contributions (central limit theorem); the "bell curve" of statistics.
Distribution. The density is

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big);$$

the distribution function has no elementary closed form.
Expectation and variance. We first verify that the standardisation $Z := \frac{X-\mu}{\sigma}$ is standard normal (here $\sigma > 0$). For every $z \in \mathbb{R}$, the change of variables $x = \mu + \sigma u$ (so $dx = \sigma\, du$) gives

$$\mathbf{P}(Z \le z) = \mathbf{P}(X \le \mu + \sigma z) = \int_{-\infty}^{\mu+\sigma z} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big)\, dx = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-u^2/2}\, du,$$

so Z has density $\frac{1}{\sqrt{2\pi}} e^{-z^2/2}$, that is $Z \sim N(0,1)$. For the standard normal, by symmetry and integration by parts,

$$\mathbf{E}[Z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x\, e^{-x^2/2}\, dx = 0, \qquad \mathbf{Var}[Z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\, dx = 1.$$

Therefore

$$\mathbf{E}[X] = \sigma\, \mathbf{E}[Z] + \mu = \mu, \qquad \mathbf{Var}[X] = \sigma^2\, \mathbf{Var}[Z] = \sigma^2.$$

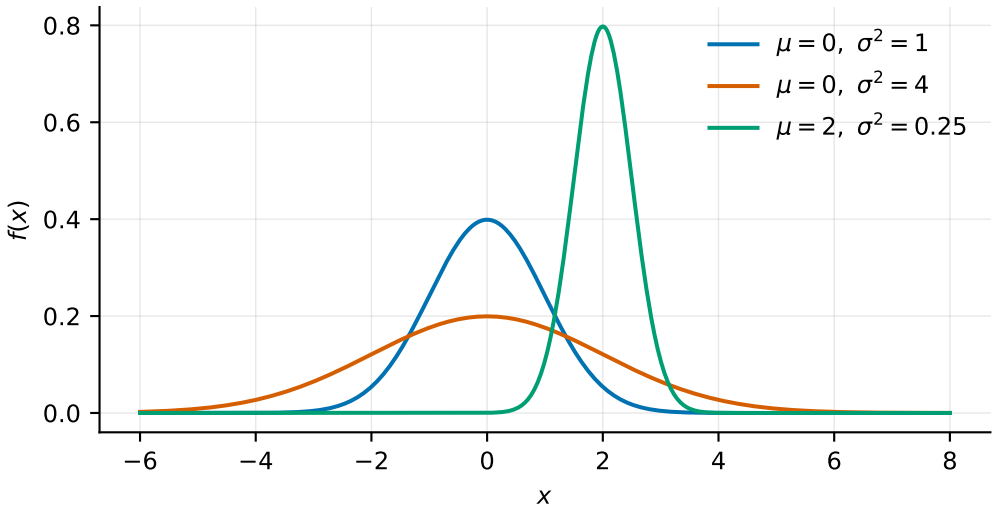

Figure A.9. Density of $N(\mu, \sigma^2)$ for three choices of mean and variance.

A.12. **Gamma distribution $\Gamma(\alpha, \lambda)$. Experiment.** For integer α, the waiting time until the α-th event when events arrive after independent $\exp(\lambda)$-distributed times; i.e. the sum of α independent exponential waiting times (the continuous analogue of the negative binomial).
Distribution. For parameters $\alpha, \lambda > 0$, the density is

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \qquad x \geq 0, \qquad \Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt;$$

here $\Gamma(\alpha) = (\alpha - 1)!$ for integer α. The case $\alpha = 1$ is $\exp(\lambda)$.
Expectation and variance. For integer α, write $X = Y_1 + \cdots + Y_\alpha$ with independent $Y_i \sim \exp(\lambda)$, so that

$$\mathbf{E}[X] = \frac{\alpha}{\lambda}, \qquad \mathbf{Var}[X] = \frac{\alpha}{\lambda^2}.$$

In general, using $\int_0^\infty x^{\alpha-1} e^{-\lambda x}\, dx = \Gamma(\alpha)/\lambda^\alpha$,

$$\mathbf{E}[X] = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^\alpha e^{-\lambda x}\, dx = \frac{\Gamma(\alpha+1)}{\lambda\, \Gamma(\alpha)} = \frac{\alpha}{\lambda}, \qquad \mathbf{E}[X^2] = \frac{\Gamma(\alpha+2)}{\lambda^2\, \Gamma(\alpha)} = \frac{\alpha(\alpha+1)}{\lambda^2},$$

so

$$\mathbf{Var}[X] = \frac{\alpha(\alpha+1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}.$$

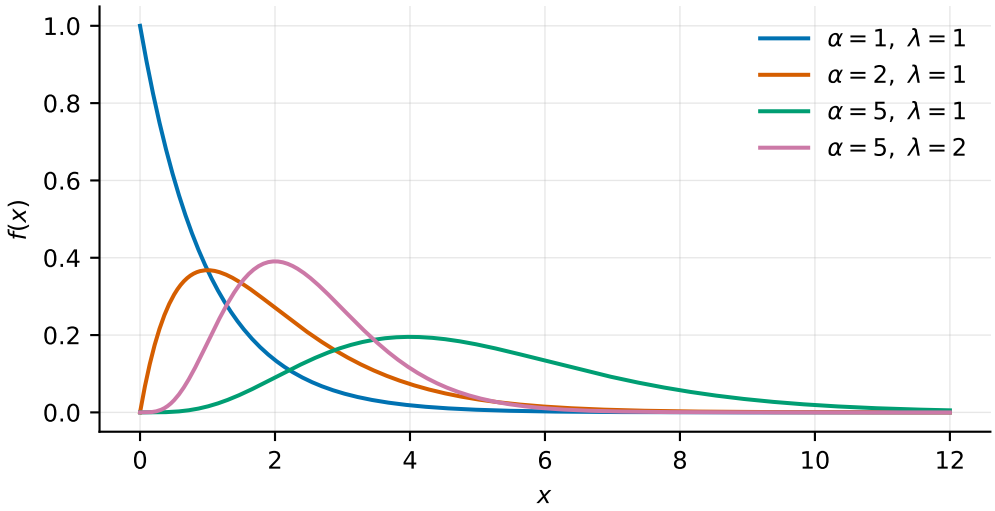

Figure A.10. Density of $\Gamma(\alpha, \lambda)$ for several shape parameters α and rates λ; $\alpha = 1$ is $\exp(\lambda)$.

Note on the use of AI. This manuscript was prepared with the help of an AI assistant (Claude Opus 4.8). It helped merge two earlier manuscripts, polished the exposition, produced the distribution plots in the appendix, and checked the final text for consistency. Responsibility for all content, and for any remaining errors, rests with the author.