1

Probability and Randomness

We start with two motivating examples of stochastic models that arise in scientific applications. The purpose of these examples is to pose scientifically meaningful questions about the properties of these models, whetting the reader's appetite for the upcoming mathematical and statistical tools that will help us answering these questions. We review some probability concepts that will prove useful later. Random behavior of simple or complex phenomena can sometimes be explained in physical terms, with an additional touch of probability theory. We exemplify this with a description of coin tossing. Finally, we define a stochastic process, give some examples of uses of stochastic models, and provide an overview of the book.

1.1 Motivating Examples

Example: *Wright-Fisher model of genetic drift* Consider a population of m individuals. All individuals are diploid, meaning that each member of the population carries two copies of her chromosomes. Suppose we are interested in a particular "gene" on one of the chromosomes. Here, a "gene" is loosely defined as stretch of DNA, not necessarily a protein coding region. Since our population is diploid, we have $2m$ genes in total. Suppose that the gene under consideration has only two possible variants, called alleles in genetics: A and a. Let X_n be the number of A alleles in the population at generation n. We assume that there is no selection acting on our gene, so we are interested in random fluctuations of allele frequencies under neutrality. We also assume that the population size stays constant and that individuals in the population mate at random with each other. Under all these assumptions, it is reasonable to postulate the following stochastic mechanism for changes in allele frequencies (counts). During each reproductive cycle, the genes of the next generation are obtained by sampling with replacement genes from the previous generation. To keep the population size constant, the number of samples is equal to $2m$. This is the simplest formulation of the Wright-Fisher model in population genetics (Fisher, 1930; Wright, 1931). Using tools from stochastic modeling, we would like to answer the following questions:

1. By construction, X_n will eventually get absorbed into either $X_\infty = 0$ or $X_\infty = 2m$. What is the probability that X_n get absorbed in $X_\infty = 0$ vs. $X_\infty = 2m$?

2. How quickly does the absorption occur?

3. How can we extend this model to achieve non-trivial allele frequencies in the population at equilibrium?

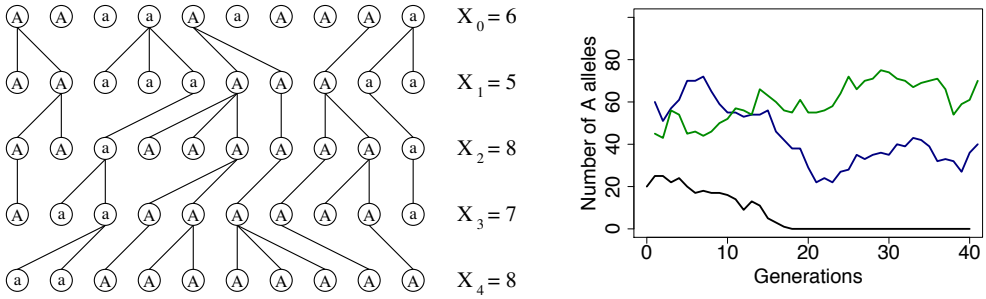

FIGURE 1.1: Wright-Fisher model of genetic drift. The left plot shows five generations of the Wright-Fisher process in a population of 10 individuals. Each row represents alleles (circles) in a particular generation. Lines connecting circles denote sampling with replacement. The number of A alleles for each generation is shown next to each row. The right plot show three realizations of the Wright-Fisher process, tracking the number of A for 40 generations.

Ronald Fisher
1890 – 1962

British statistician and geneticist Ronald Fisher and American geneticist Sewall Wright, together with a British mathematical biologist, Jack Haldane, laid a foundation for what is now known as modern evolutionary synthesis – a theory that reconciled seeming discrepancies between Darwinian and Mendelian schools of thought on evolution and natural selection. The Wright-Fisher model and its various modifications played a pivotal role in this synthesis.

Sewall Wright
1889 – 1988

Example: *Susceptible-Infected-Recovered (SIR) epidemic model* Consider a population of N individuals. Let I_t be the number of individuals infected with an disease (e.g., influenza), S_t be the number of susceptible individuals, and R_t be the number of recovered or removed invividuals. Assuming that $I_0 = 1$, $S_0 = N-1$, and $R_t = 0$, we would like to model the state of the epidemic, (I_t, S_t, R_t) as a continuous-time Markov chain. We further assume that recovered individuals stay recovered forever and that $I_t + S_t + R_t = N$.

We are interested in answering the following questions:

1. How to define this Markov process so it can produce reasonable stochastic behavior of an epidemic in a closed population? This model is called an SIR model in infectious disease epidemiology.

2. What is the distribution of the total number of individuals that become infected during the course of the epidemic?

3. What is the probability that all individuals in the population become infected?

4. How can we infer parameters of the SIR model given partial information about infected/uninfected status of individuals in the populations?

5. How do we make forecasts about progression of the epidemic given noisy data during the initial stage of the epidemic?

Notice that the last two questions usually fall outside of the scope of a typical book on applied

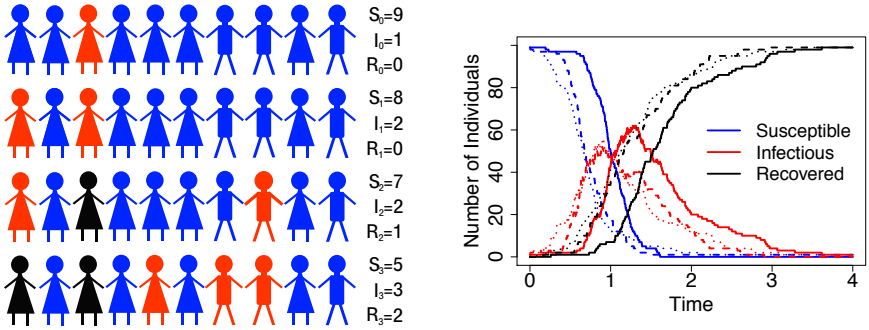

FIGURE 1.2: Susceptible-infectious-recovered (SIR) stochastic epidemic model. The left plot shows 4 times points during an epidemic in the population with 10 individuals. Susceptible individuals are shown in blue, infectious individuals are shown in red, and recovered individuals are shown in black. The right plot shows three realizations of the SIR epidemic model.

probability and/or stochastic processes. Our exposition of stochastic processes is different in that we are primarily interested in applications of stochastic processes to data analyses arising in sciences.

Anderson Gray McKendrick 1875 – 1943

In 1766 the Swiss mathematician and physicist Daniel Bernolli used mathematics to argue for the benefits of vaccination against smallpox. Anderson Gray McKendrick was a Scottish army officer, director of the Pasteur Institute 1920-1926, and after that superintendent of the Royal College of Physicians' laboratory, Edinburgh. In 1914 he published the first account of a pure birth process. His 1926 DSc dissertation at Aberdeen was entitled "Applications of mathematics to medical problems" and may be the first stochastic model of an epidemic (although Bailey (1986) mentions a probabilistic model by an Estonian physician named En'ko in 1889). The modern stochastic epidemic theory was essentially created by Norman T. J. Bailey, an English statistician who spent most of his career at the World Health Organization in Geneva.

Norman T.J. Bailey 1923 –

For readers that need to brush up on concepts and standard notation used in probability theory, we provide a probability refresher in the next section.

1.2 Probability Refresher

We assume that we can assign probabilities to *events* — outcomes of a random experiment. For example, tossing a coin results in one of two possible events: H ="heads" and T="tails." More formally, we define *events* as certain subsets of some abstract space.

Definition. The *state space*[1] Ω is a collection of all possible outcomes of a random experiment.

[1]We use the term *state space*, which is frequently used in stochastic processes theory. In probability textbooks the term *sample space* is more common.

Next, we need to be able to assign probabilities to events. This is done by introducing a probability measure.

Definition. A *probability measure* $\Pr()$ is a function mapping subsets of Ω to real numbers, satisfying:

1. $\Pr(\Omega) = 1$.
2. $0 \leq \Pr(A) \leq 1$ $A \subset \Omega$.
3. $\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i)$ for *mutually exclusive* events $A_1, A_2, \ldots$

Note 1.1. In the case of countable Ω we can define Pr for all subsets, while if the cardinality is larger we have to restrict attention to certain subsets, called measuarble (Durret, 2004).

Example: *Discrete uniform distribution* Let $\Omega = \{1, 2, \ldots, n\}$, we define $\Pr(A) = |A|/n$, where $|A|$ is the number of elements in $A \in \Omega$. For example, if $n = 10$, then $\Pr(\{1\}) = 0.1$ and $\Pr(\{2, 9, 10\}) = 0.3$.

We also need a concept of a random variable. Informally, a random variable X is a function or variable, whose value is generated by a random experiment. For example, we can define a binary random variable associated with a toss of a coin:

$$X = \begin{cases} 1 & \text{if heads,} \\ 0 & \text{if tails.} \end{cases}$$

The formal definition is below.

Definition. A function $X(\omega) : \Omega \to \bar{\mathbb{R}}$ is called a *random variable* (r.v.).

Note 1.2. Later in the book, we will see random variables satisfying $\Pr(X = \infty) > 0$. Hence, we map Ω to $\bar{\mathbb{R}} = \mathbb{R} \bigcup \{\pm\infty\}$.

Definition. For events A and B in Ω we define *conditional probability*

$$\Pr(B \,|\, A) = \frac{\Pr(A \bigcap B)}{\Pr(A)}.$$

If $\Pr(B \,|\, A) = \Pr(B)$ we say that the events A and B are *independent*, which together with the formula above implies that $\Pr(A \bigcap B) = \Pr(A) \times \Pr(B)$. Two r.v.s X and Y are called independent if the events $\{X \in A\}$ and $\{Y \in B\}$ are independent for all A and B. A sequence $X_1, \ldots X_n$ of random variables is called iid (independent and identically distributed) if they are mutually independent and have the same distribution.

Example: *Bernoulli r.v.* $X \in \{0, 1\}$ with $\Pr(X = 1) = p$, $\Pr(X = 0) = 1 - p$ for $0 \leq p \leq 1$ is called a *Bernoulli* random variable with parameter (or success probability) p.

Example: *Binomial r.v.* Let $X_i \sim \text{Bernoulli}(p)$ be independent. Then the number of successes $S_n = \sum_{i=1}^{n} X_i$ is called a *binomial r.v.* with

$$\Pr(S_n = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

Example: *Geometric r.v.* If $X_1, X_2, \ldots$ are independent Bernoulli(p), let $N = \min\{n : X_n = 1\}$ be the number of trials until the first success occurs, including the successful trial. Then

$$\Pr(N = n) = (1-p)^{n-1} p \text{ for } n = 1, 2, \ldots$$

Note 1.3. There is an alternative definition of the geometric distribution does not count the successful trial so that $\Pr(N = n) = (1-p)^n p$ for $n = 0, 1, \ldots$.

We defined all discrete random variables above using probabilities of X taking a particular value. A function that assigns probabilities to random variable values is called a *probability mass function*. However, a more general way to define random variables is by specifying a *cumulative distribution function*.

Definition. $F(x) = \Pr(X \leq x)$ is called the *cumulative distribution function* (cdf) of X.

Properties of cdf:

1. $0 \leq F(x) \leq 1$.
2. $F(x) \leq F(y)$ for $x \leq y$.
3. $\lim_{x \to y^+} F(x) = F(y)$ ($F(x)$ is right-continuous).
4. $\lim_{x \to -\infty} F(x) = \Pr(X = -\infty)$ (usually $= 0$).
5. $\lim_{x \to \infty} F(x) = 1 - \Pr(X = \infty)$ (usually $= 1$).
6. $\Pr(X = x) = F(x) - F(x-)$ where $F(x-) = \lim_{y \uparrow x} F(y)$.

Example: *Discrete uniform random variable* For a random variable U, uniformly distributed over $\{1, 2, \ldots, n\}$, its cdf is given by

$$F(x) = \begin{cases} 0 & \text{if } x < 1, \\ 1/n & \text{if } 1 \leq x < 2, \\ 2/n & \text{if } 2 \leq x < 3, \\ \vdots & \\ (n-1)/n & \text{if } n-1 \leq x < n, \\ 1 & \text{if } x \geq n. \end{cases}$$

The probability mass function and cdf of U, with $n = 10$, are shown in Figure 1.3, which also contains the probability mass function and cdf of a geometric random variable.

For continuous random variables, the analog of the probability mass function is a probability density function, defined as follows.

Definition. If $F(x) = \int_{-\infty}^{x} f(x)dx$ for some $f(x) \geq 0$, then $f(x)$ is called *probability density function* of X. If X has a probability density function, we say that X is *absolutely continuous*.

Note 1.4. $\int_a^b f(x)\mathrm{d}x = F(b) - F(a) = \Pr(a \leq X \leq b)$ for $a \leq b$. Moreover, $\frac{d}{dx} F(x) = f(x)$.

Example: *Uniform random variable on* $[0, 1]$ The random variable U with density

$$f(x) = \begin{cases} 1 & \text{if } x \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}$$

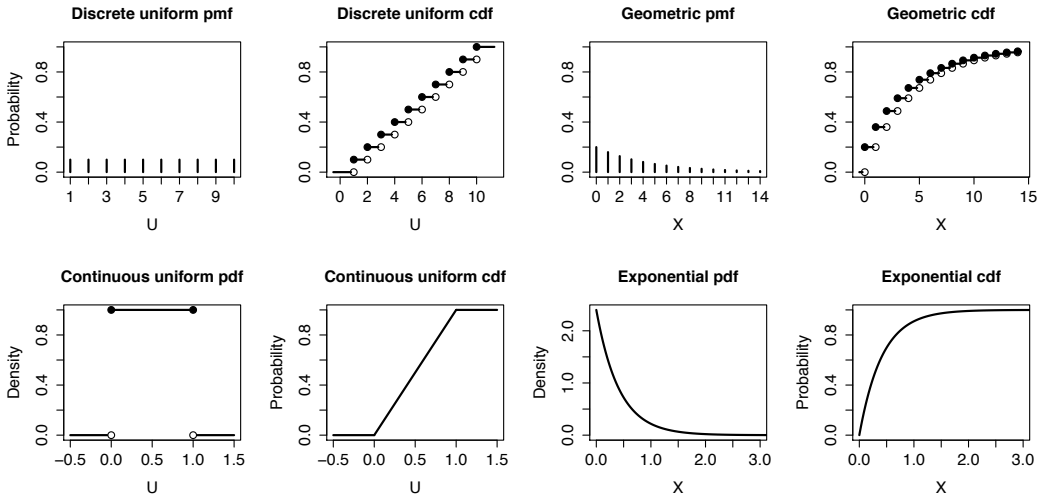

FIGURE 1.3: Probability mass functions (pmfs) and cumulative distribution functions (cdfs) for the discrete uniform random variable over $\{1,2,\ldots,10\}$ and for the geometric random variable with success probability $p = 0.2$ (top row). Probability density functions (pdfs) and cdfs for the continuous uniform random variable over $[0,1]$ and for the exponential random variable with rate parameter $\lambda = 2.4$ (bottom row).

is called *uniform*. The cdf of U is

$$F(x) = \begin{cases} 0 & \text{if } x < 0. \\ x & \text{if } 0 \le x \le 1, \\ 1 & \text{if } x > 1. \end{cases}$$

Figure 1.3 shows the above functions below their discrete analogs.

Now that we know how to define univariate random variables using pmfs/pdfs or cdfs, we need to develop a vocabulary for describing properties of these distribution. The concept of expectation is central to this development.

Definition. The *expectation* of $g(X)$ is defined as $\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x)dF(x)$, where the integral is taken (in the Stieltjes sense) with respect to the probability measure. More concretely,

1. For a discrete random variable X, $\mathrm{E}[g(X)] = \sum_{k=1}^{\infty} g(x_k)\Pr(X = x_k)$.
2. For an absolutely continuous random variable X, $\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$.

The two most important expectations are $\mathrm{E}(X)$, called the *mean* of the r.v., and $\mathrm{Var}(X) = \mathrm{E}\left\{[X - E(X)]^2\right\}$, called the *variance* of the r.v.. We often describe a random variable and its distribution by referring to their mean and variance.

Example: *Exponential r.v.* An exponential random variable has density $f(x) = \lambda e^{-\lambda x} 1_{\{x \ge 0\}}$, where $\lambda > 0$ is the rate parameter. Let $X \sim \mathrm{Exp}(\lambda)$. Then

$$\mathrm{E}(X) = \int_0^{\infty} x\lambda e^{-\lambda x}dx = \begin{bmatrix} u = x & e^{-\lambda x}dx = dv \\ du = dx & -\frac{e^{-\lambda x}}{\lambda} = v \end{bmatrix} = \lambda \left[-x\frac{e^{-\lambda x}}{\lambda}\Big|_0^{\infty} + \int_0^{\infty} \frac{e^{-\lambda x}}{\lambda}dx \right] = \lambda \left[0 + \frac{1}{\lambda^2} \right] = \frac{1}{\lambda}.$$

Expectations are linear operators, meaning that for any collection of random variables $X_1,\ldots,X_n$ and real numbers $a_1,\ldots a_n$

$$\mathrm{E}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i \mathrm{E}(X_i).$$

Linearity does not hold for the variance in general. However, if random variables $X_1,\ldots,X_n$ are independent, then

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i).$$

Definition. If we have a r.v. X defined on Ω, then we can define *conditional expectation*

$$\mathrm{E}(X\,|\,A) = \frac{E(X 1_{\{A\}})}{\Pr(A)}.$$

Conditioning on random variables can be a little tricky, so we will limit our discussion of this concept to the

1. discrete case:
$$\Pr(X = x\,|\,Y = y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)},$$
and the
2. continuous case:
$$F_{X|Y}(x|y) = \frac{\int_{-\infty}^{x} f_{XY}(z,y)dz}{f_Y(y)} \qquad \text{and} \qquad f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)},$$
where $f_{XY}(x,y)$ is the joint density of X and Y and $f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx$ is the marginal density of Y.

Note 1.5. If r.v.s X and Y are independent, then $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$ and $\mathrm{E}(X\,|\,Y) = \mathrm{E}(X)$. The last equality says that Y carries no information about the location of X.

Example: *Hypergeometric distribution* Let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$, $S_n = \sum_{i=1}^{n} X_i$, and $S_m = \sum_{i=1}^{m} X_i$ for $m < n$. We want to find the distribution of S_m conditional on S_n. We start with probability mass function

$$\Pr(S_m = j\,|\,S_n = k) = \frac{\Pr(S_m = j, S_n = k)}{\Pr(S_n = k)} = \frac{\Pr(\sum_{i=1}^{m} X_i = j, \sum_{i=1}^{n} X_i = k)}{\Pr(S_n = k)}$$
$$= \frac{\Pr\left(\sum_{i=1}^{m} X_i = j, \sum_{i=m+1}^{n} X_i = k-j\right)}{\Pr(S_n = k)} = [\text{independence}] = \frac{\Pr(\sum_{i=1}^{m} X_i = j)\Pr\left(\sum_{i=m+1}^{n} X_i = k-j\right)}{\Pr(S_n = k)}$$
$$= \frac{\binom{m}{j} p^j (1-p)^{m-j} \binom{n-m}{k-j} p^{k-j}(1-p)^{n-m-k+j}}{\binom{n}{k} p^k (1-p)^{n-k}} = \frac{\binom{m}{j}\binom{n-m}{k-j}}{\binom{n}{k}}.$$

This is the probability mass function of the *hypergeometric* distribution.

$$\mathrm{E}(S_m\,|\,S_n = k) = \sum_{i=1}^{m} \mathrm{E}(X_i\,|\,S_n = k) = [\text{symmetry}] = m\mathrm{E}(X_1\,|\,S_n = k) = \frac{m}{n}\sum_{i=1}^{n} \mathrm{E}(X_i\,|\,S_n = k)$$
$$= \frac{m}{n}\mathrm{E}(S_n\,|\,S_n = k) = \frac{mk}{n}.$$

Notice that $X_1,\ldots,X_n$ do not have to be Bernoulli for $\mathrm{E}(S_m\,|\,S_n) = mS_n/n$ to hold.

Law of total probability

If $B_1, \ldots, B_n$ are mutually exclusive events and $\bigcup_{i=1}^{n} B_i = \Omega$, then

$$\Pr(A) = \sum_{i=1}^{n} \Pr\left(A \bigcap B_i\right) = \sum_{i=1}^{n} \Pr(A \,|\, B_i) \Pr(B_i).$$

Law of total expectation

Recall that $\mathrm{E}(X)$ is a scalar, but $\mathrm{E}(X \,|\, Y)$ is a random variable. Let X and Y be discrete r.v.s. and define

$$\mathrm{E}(X \,|\, Y = y) = \sum_{k=1}^{\infty} x_k \Pr(X = x_k \,|\, Y = y).$$

Then $\mathrm{E}[\mathrm{E}(X \,|\, Y)] = \mathrm{E}(X)$.

Proof.

$$\mathrm{E}[\mathrm{E}(X \,|\, Y)] = \sum_{k=1}^{\infty} \mathrm{E}(X \,|\, Y = y_k) \Pr(Y = y_k) = \sum_{k=1}^{\infty} \frac{\mathrm{E}\left(X 1_{\{Y=y_k\}}\right)}{\Pr(Y = y_k)} \Pr(Y = y_k)$$
$$= \sum_{k=1}^{\infty} \mathrm{E}\left(X 1_{\{Y=y_k\}}\right) = \mathrm{E}\left(X 1_{\{\bigcup_{k=1}^{\infty}\{Y=y_k\}\}}\right) = \mathrm{E}(X).$$

□

As long as the quantities are defined, $\mathrm{E}[\mathrm{E}(X \,|\, Y)] = \mathrm{E}(X)$. In fact, this equality is often used as a definition of the conditional expectation, when conditioning on a random variable (Durret, 2004).

Law of total variance

Decomposing variance using conditioning is only slightly more complicated:

$$\mathrm{Var}(X) = \mathrm{E}\left(X^2\right) - \mathrm{E}(X)^2 = [\text{law of total expectation}] = \mathrm{E}\left[\mathrm{E}\left(X^2 \,|\, Y\right)\right] - \mathrm{E}[\mathrm{E}(X \,|\, Y)]^2$$
$$= [\text{def of variance}] = \mathrm{E}\left[\mathrm{Var}(X \mid Y) + \mathrm{E}(X \,|\, Y)^2\right] - \mathrm{E}[\mathrm{E}(X \,|\, Y)]^2 = \mathrm{E}[\mathrm{Var}(X \mid Y)]$$
$$+ \left\{\mathrm{E}\left[\mathrm{E}(X \,|\, Y)^2\right] - \mathrm{E}[\mathrm{E}(X \,|\, Y)]^2\right\} = [\text{def of variance}] = \mathrm{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}[\mathrm{E}(X \mid Y)].$$

Right tail probability trick for taking expectations

In Markov chain theory, we will often deal with non-negative random variables, whose expectations can be expressed in terms of their cdfs.

Discrete r.v. on non-negative integers

Let $X \in \{0, 1, \ldots\}$.

$$\mathrm{E}(X) = \sum_{k=0}^{\infty} k \Pr(X = k) = \sum_{k=1}^{\infty} k \Pr(X = k) =$$
$$\Pr(X = 1) +$$
$$\Pr(X = 2) + \Pr(X = 2) +$$
$$\Pr(X = 3) + \Pr(X = 3) + \Pr(X = 3) +$$
$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

$$[\text{summing up columns}] = \sum_{k=1}^{\infty} \Pr(X=k) + \sum_{k=2}^{\infty} \Pr(X=k) + \sum_{k=3}^{\infty} \Pr(X=k) + \dots$$
$$= \Pr(X \geq 1) + \Pr(X \geq 2) + \Pr(X \geq 3) + \cdots = \sum_{k=1}^{\infty} \Pr(X \geq k).$$

Non-negative absolutely continuous r.v.

Let $F(x)$ be the cdf of a non-negative r.v. X. Assuming that $\int_0^\infty [1-F(x)]dx < \infty$ write

$$\mathrm{E}(X) = \int_o^\infty t f(t) dt = \begin{bmatrix} u = t & f(t)dt = dv \\ du = dt & -[1-F(t)] = v \end{bmatrix} = -t[1-F(t)]\Big|_0^\infty + \int_0^\infty [1-F(t)]dt$$
$$= \int_0^\infty [1-F(t)]dt = \int_0^\infty \Pr(X > t)\, dt.$$

Note 1.6. We used $\lim_{t\to\infty} t[1-F(t)] = 0$, which follows from the fact that $1-F(t)$ is non-increasing and our assumption $\int_0^\infty [1-F(x)]dx < \infty$. Intuitively, in order for the last integral to be finite, $1-F(t)$ must decrease to 0 faster than $1/t$.

Note 1.7. A more general formulation of the above results, available for example in (Lange, 2004, Chapter 2), states that for any integrable $h(x)$ and $H(x) = H(0) + \int_0^x h(t)dt$,

$$\mathrm{E}[H(X)] = H(0) + \int_0^\infty h(t)[1-F(t)]dt.$$

Probability limits

Definition. Almost sure convergence

Definition. Convergence in distribution

Limit theorems

Theorem 1.1. *Strong Law of Large Numbers (SLLN). Let $X_1, X_2, \dots$ be independent and identically distributed random variables with $\mu = E(X_1) < \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mu.$$

SLLN says that the empirical average of iid random variables converges to the theoretical average/expectation.

Theorem 1.2. *Central Limit Theorem (CLT). Let $X_1, X_2, \dots$ be independent and identically distributed random variables with $\mu = E(X_1) < \infty$ and $0 < \sigma^2 = Var(X_1) < \infty$ and let $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Then*

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim \mathcal{N}(0,1) \text{ approximately for large } n.$$

Informally, the CLT says that for large n, the empirical average behaves as $\mathcal{N}(\mu, \sigma^2/n)$. Scaling of the variance by $1/n$ implies that averaging reduces variability, which makes intuitive sense.

1.3 Randomness

The world is full of unpredictable events. Science strives to understand natural phenomena, in the sense of reducing this unpredictability. There are many ways of doing this. Models, which are ab-