Chapter 2

A Quick Review of Probability Theory

These notes are not intended to be a comprehensive introduction to the theory of probability. Instead, they constitute a brief introduction that should be sufficient to allow a student to understand the stochastic models they will encounter in later chapters. These notes were heavily influenced by Sheldon Ross's text [14], and Timo Seppäläinen's notes on probability theory that serve a similar purpose [15]. Any student who finds this material difficult should review an introductory probability book such as Sheldon Ross's *A first course in probability* [14], which is on reserve in the Math library.

2.1 The Probability Space

Probability theory is used to model experiments (defined loosely) whose outcome can not be predicted with certainty beforehand. For any such experiment, there is a triple $(\Omega, \mathcal{F}, P)$, called a *probability space*, where

- Ω is the *sample space*,
- $\mathcal{F}$ is a collection of *events*,
- P is a *probability measure*.

We will consider each in turn.

2.1.1 The sample space Ω

The *sample space* of an experiment is the set of all possible outcomes. Elements of Ω are called *sample points* and are often denoted by ω. Subsets of Ω are referred to as *events*.

Example 2.1.1. Consider the experiment of rolling a six-sided die. Then the natural sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. □

Example 2.1.2. Consider the experiment of tossing a coin three times. Let us write 1 for heads and 0 for tails. Then the sample space consists of all sequences of length three consisting only of zeros and ones. Each of the following representations is valid

$$\begin{aligned}\Omega &= \{0,1\}^3 \\ &= \{0,1\} \times \{0,1\} \times \{0,1\} \\ &= \{(x_1, x_2, x_3) : x_i \in \{0,1\} \text{ for } i = 1,2,3\} \\ &= \{(0,0,0), (0,0,1), (0,1,0), (1,0,0), (0,1,1), (1,0,1), (1,1,0), (1,1,1)\}.\end{aligned}$$

□

Example 2.1.3. Consider the experiment of counting the number of mRNA molecules transcribed by a given gene in some interval of time. Here it is most natural to let $\Omega = \{0, 1, 2, \dots\}$. □

Example 2.1.4. Consider the experiment of waiting for a bacteria to divide. In this case, it is natural to take as our sample space all values greater than or equal to zero. That is, $\Omega = \{t \ : \ t \geq 0\}$, where the units of t are specified as hours, for example.□

Note that the above sample spaces are quite different in nature. Those of Examples 2.1.1 and 2.1.2 are finite, while those of 2.1.3 and 2.1.4 are infinite. The sample space of Example 2.1.3 is countably infinite while that of Example 2.1.4 is uncountably infinite. A set that is finite or countably infinite is called *discrete*. Most, though not all, of the sample spaces encountered in this course will be discrete, in which case probability theory requires no mathematics beyond calculus and linear algebra.

2.1.2 The collection of events $\mathcal{F}$

Events are simply subsets of the state space Ω. They are often denoted by A, B, C, etc., and they are usually the objects we wish to know the probability of. They can be described in words, or using mathematical notation. Examples of events of the experiments described above are the following:

Example 2.1.1, continued. Let A be the event that a 2 or a 4 is rolled. That is, $A = \{2, 4\}$. □

Example 2.1.2, continued. Let A be the event that the final two tosses of the coin are tails. Thus,

$$A = \{(1,0,0), (0,0,0)\}.$$

□

Example 2.1.3, continued. Let A be the event that no more than 10 mRNA molecules have appeared. Thus,

$$A = \{0, 1, 2, \dots, 10\}.$$

□

Example 2.1.4, continued. Let A be the event that it took longer than 2 hours for the bacteria to divide. Then,

$$A = \{t \ : \ t > 2\}.$$

□

We will often have need to consider the unions and intersections of events. We write $A \cup B$ for the union of A and B, and either $A \cap B$ or AB for the intersection.

For discrete sample spaces, $\mathcal{F}$ will contain all subsets of Ω, and will play very little role. This is the case for nearly all of the models in this course. When the state space is more complicated, $\mathcal{F}$ is assumed to be a σ-algebra. That is, it satisfies the following three axioms:

1. $\Omega \in \mathcal{F}$.
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, where A^c is the complement of A.
3. If $A_1, A_2, \ldots, \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

2.1.3 The probability measure P

Definition 2.1.5. The real valued function P, with domain $\mathcal{F}$, is a *probability measure* if it satisfies the following three axioms

1. $P(\Omega) = 1$.
2. If $A \in \mathcal{F}$ (or equivalently for discrete spaces if $A \subset \Omega$), then $P(A) \geq 0$.
3. If for a sequence of events $A_1, A_2, \ldots,$, we have that $A_i \cap A_j = \emptyset$ for all $i \neq j$ (i.e. the sets are *mutually exclusive*) then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

The following is a listing of some of the basic properties of any probability measure, which are stated without proof.

Lemma 2.1.6. *Let $P(\cdot)$ be a probability measure. Then*

1. If $A_1, \ldots, A_n$ is a finite sequence of mutually exclusive events, then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

2. $P(A^c) = 1 - P(A)$.

3. $P(\emptyset) = 0$.

4. *If* $A \subset B$, *then* $P(A) \leq P(B)$.

5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Note that for discrete spaces, we can (at least theoretically) find $P(A)$ for any $A \in \mathcal{F}$ so long as we know $P(\omega)$ for every $\omega \in \Omega$.

Example 2.1.7. Suppose we roll an unfair die that yields a 1, 2, or 3 each with a probability of 1/10, that yields a 4 with a probability of 1/5, and yields a 5 or 6 each with a probability of 1/4. Then, the probability we roll an even number is

$$P\{2,4,6\} = P\{2\} + P\{4\} + P\{6\} = \frac{1}{10} + \frac{1}{5} + \frac{1}{4} = \frac{11}{20}.$$

□

2.2 Conditional Probability and Independence

Suppose we are interested in the probability that some event A took place, though we have some extra information in that we know some other event B took place. For example, suppose that we want to know the probability that a fair die rolled a 4 given that we know an even number came up. Most people would answer this as 1/3, as there are three possibilities for an even number, $\{2,4,6\}$, and as the die was fair, each of the options should be equally probable. The following definition generalizes this intuition.

Definition 2.2.1. For two events $A, B \subset \Omega$, the *conditional probability of A given B* is

$$P(A|B) = \frac{P(AB)}{P(B)}.$$

provided that $P(B) > 0$.

Example 2.2.2. The probability that it takes a bacteria over 2 hours to divide is 0.64, and the probability it takes over three hours is 0.51. What is the probability that it will take over three hours to divide, given that two hours have already passed?

Solution: Let A be the event that the bacteria takes over three hours to split and let B be the event that it takes over two hours to split. Then, because $A \subset B$,

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)}{P(B)} = \frac{.51}{.64} \approx 0.797.$$

□

We intuitively think of A being independent from B if $P(A|B) = P(A)$, and $P(B|A) = P(B)$. More generally, we have the following definition.

Definition 2.2.3. The events $A, B \in \mathcal{F}$ are called *independent* if

$$P(AB) = P(A)P(B).$$

It is easy to check that the definition of independence implies both $P(A|B) = P(A)$ and $P(B|A) = P(B)$, and vice versa when $P(A) > 0$ and $P(B) > 0$. The concept of independence will play a key role in our study of Markov chains.

Theorem 2.2.4. *Let Ω be a sample space with $B \in \mathcal{F}$, and $P(B) > 0$. Then*

(a) $P(A \mid B) \geq 0$ for any event $A \in \mathcal{F}$.

(b) $P(\Omega \mid B) = 1$.

(c) If $A_1, A_2, \cdots \in \mathcal{F}$ is a sequence of mutually exclusive events, then

$$P\left(\bigcup_{i=1}^{\infty} A_i \mid B\right) = \sum_{i=1}^{\infty} P(A_i|B).$$

Therefore, conditional probability measures are themselves probability measures, and we may write $Q(\cdot) = P(\cdot \mid B)$.

By definition

$$P(A|B) = \frac{P(AB)}{P(B)}, \quad \text{and} \quad P(B|A) = \frac{P(AB)}{P(A)}.$$

Rearranging terms yields

$$P(AB) = P(A|B)P(B), \quad \text{and} \quad P(AB) = P(B|A)P(A).$$

This can be generalized further to the following.

Theorem 2.2.5. *If $P(A_1A_2 \cdots A_n) > 0$, then*

$$P(A_1A_2 \cdots A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1A_2)P(A_4 \mid A_1A_2A_3) \cdots P(A_n \mid A_1A_2 \cdots A_{n-1}).$$

Definition 2.2.6. Let $\{B_1, \ldots, B_n\}$ be a set of nonempty subsets of $\mathcal{F}$. If the sets B_i are mutually exclusive and $\bigcup B_i = \Omega$, then the set $\{B_1, \ldots, B_n\}$ is called a *partition* of Ω.

Theorem 2.2.7 (Law of total probability)**.** *Let $\{B_1, \ldots, B_n\}$ be a partition of Ω with $P(B_i) > 0$. Then for any $A \in \mathcal{F}$*

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

Theorem 2.2.8 (Bayes' Theorem)**.** *For all events $A, B \in \mathcal{F}$ such that $P(B) > 0$ we have*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

This allows us to "turn conditional probabilities around."

Proof. We have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

□

2.3 Random Variables

Henceforth, we assume the existence of some probability space $(\Omega, \mathcal{F}, P)$. All random variables are assumed to be defined on this space in the following manner.

Definition 2.3.1. A *random variable* X is a real-valued function defined on the sample space Ω. That is, $X : \Omega \to \mathbb{R}$.

If the range of X is finite or countably infinite, then X is said to be a *discrete random variable*, whereas if the range is an interval of the real line (or some other uncountably infinite set), then X is said to be a *continuous random variable*.

Example 2.3.2. Suppose we roll two die and take $\Omega = \{(i, j) \mid i, j \in \{1, \ldots, 6\}\}$. We let $X(i, j) = i + j$ be the discrete random variable giving the sum of the rolls. The range is $\{2, \ldots, 12\}$. □

Example 2.3.3. Consider two bacteria, labeled 1 and 2. Let T_1 and T_2 denote the times they will divide to give birth to daughter cells, respectively. Then, $\Omega = \{(T_1, T_2) \mid T_1, T_2 \geq 0\}$. Let X be the continuous random variable giving the time of the first division: $X(T_1, T_2) = \min\{T_1, T_2\}$. The range of X is $t \in \mathbb{R}_{\geq 0}$. □

Notation: As is traditional, we will write $\{X \in I\}$ as opposed to the cumbersome $\{\omega \in \Omega \mid X(\omega) \in I\}$.

Definition 2.3.4. If X is a random variable, then the function F_X, or simply F, defined on $(-\infty, \infty)$ by

$$F_X(t) = P\{X \leq t\}$$

is called the *distribution function*, or *cumulative distribution function*, of X.

Theorem 2.3.5 (Properties of the distribution function)**.** *Let X be a random variable defined on some probability space $(\Omega, \mathcal{F}, P)$, with distribution function F. Then,*

1. *F is nondecreasing. Thus, if $s \leq t$, then $F(s) = P\{X \leq s\} \leq P\{X \leq t\} = F(t)$.*

2. *$\lim_{t\to\infty} F(t) = 1$.*

3. *$\lim_{t\to-\infty} F(t) = 0$.*

4. *F is right continuous. So, $\lim_{h\to0^+} f(t+h) = f(t)$ for all $t \in \mathbb{R}$.*

For discrete random variables, it is natural to consider a function giving the probability of each possible event, whereas for continuous random variables we need the concept of a density function.

Definition 2.3.6. Let X be a discrete random variable. Then for $x \in \mathbb{R}$, the function

$$p_X(x) = P\{X = x\}$$

is called the *probability mass function* of X.

By the axioms of probability, a probability mass function p_X satisfies

$$P\{X \in A\} = \sum_{x\in A} p_X(x).$$

Definition 2.3.7. Let X be a continuous random variable with distribution function $F(t) = P\{X \leq t\}$. Suppose that there exists a nonnegative, integrable function $f : \mathbb{R} \to [0, \infty)$, or sometimes f_X, such that

$$F(x) = \int_{-\infty}^{x} f(y)dy.$$

Then the function f is called the *probability density function* of X.

We now have that for any $A \subset \mathbb{R}$ (or, more precisely, for any $A \in \mathcal{F}$, but we are going to ignore this point),

$$P\{X \in A\} = \int_A f_X(x)dx.$$

2.3.1 Expectations of random variables

Let X be a random variable. Then, the *expected value* of X is

$$\mathbb{E}[X] = \sum_{x\in\mathcal{R}(X)} xp_X(x)$$

in the case of discrete X, and

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf_X(x)dx,$$

in the case of continuous X. The functions p_X and $f_X(x)$ above are the probability mass function and density function, respectively. The expected value of a random variable is also called its *mean* or *expectation* and is often denoted μ or μ_X.

Example 2.3.8. Consider a random variable taking values in $\{1, \ldots, n\}$ with

$$P\{X = i\} = \frac{1}{n}, \quad i \in \{1, \ldots, n\}.$$

We say that X is distributed uniformly over $\{1, \ldots, n\}$. What is the expectation?

Solution. We have

$$\mathbb{E}[X] = \sum_{i=1}^{n} iP\{X = i\} = \sum_{i=1}^{n} i\frac{1}{n} = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}.$$

□

Example 2.3.9. Consider rolling a die and letting X be the outcome. Then X is uniformly distributed on $\{1, \ldots, 6\}$. Thus, $\mathbb{E}[X] = 7/2 = 3.5$. □

Example 2.3.10. Consider the weighted die from Example 2.1.7. The expectation of the outcome is

$$1 \times \frac{1}{10} + 2 \times \frac{1}{10} + 3 \times \frac{1}{10} + 4 \times \frac{1}{5} + 5 \times \frac{1}{4} + 6 \times \frac{1}{4} = \frac{83}{20} = 4.15.$$

□

Example 2.3.11. Suppose that X is exponentially distributed with density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & , \quad x \geq 0 \\ 0 & , \quad \text{else} \end{cases},$$

where $\lambda > 0$ is a constant. In this case,

$$\mathbb{E}X = \int_0^\infty x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}. \tag{2.1}$$

□

Suppose we instead want the expectation of a function of a random variable: $g(X) = g \circ X$, which is itself a random variable. That is, $g \circ X : \Omega \to \mathbb{R}$. The following is proved in any introduction to probability book.

Theorem 2.3.12. *Let X be a random variable and let $g : \mathbb{R} \to \mathbb{R}$ be a function. Then,*

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{R}(X)} g(x)p_X(x),$$

in the case of discrete X, and

$$\mathbb{E}[g(X)] = \int_\infty^\infty g(x)f_X(x)dx,$$

in the case of continuous X. The functions p_X and $f_X(x)$ are the probability mass function and density function, respectively.

An important property of expectations is that for any random variable X, real numbers $\alpha_1, \ldots, \alpha_n$, and functions $g_1, \ldots, g_n : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[\alpha_1 g_1(X) + \cdots + \alpha_n g_n(X)] = \alpha_1 \mathbb{E}[g_1(X)] + \cdots + \alpha_n \mathbb{E}[g_n(X)].$$

2.3.2 Variance of a random variable

The variance gives a measure on the "spread" of a random variable around its mean.

Definition 2.3.13. Let μ denote the mean of a random variable X. The *variance* and *standard deviation* of X are

$$\begin{aligned}\mathsf{Var}(X) &= \mathbb{E}[(X-\mu)^2] \\ \sigma_X &= \sqrt{\mathsf{Var}(X)},\end{aligned}$$

respectively. Note that the units of the variance are the square of the units of X, whereas the units of standard deviation are the units of X and $\mathbb{E}[X]$.

A sometimes more convenient formula for the variance can be computed straightaway:

$$\mathsf{Var}(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - \mu^2.$$

A useful fact that follows directly from the definition of the variance is that for any constants a and b,

$$\begin{aligned}\mathsf{Var}(aX+b) &= a^2\mathsf{Var}(X) \\ \sigma_{aX+b} &= |a|\sigma_X.\end{aligned}$$

2.3.3 Some common discrete random variables

Bernoulli random variables: X is a *Bernoulli* random variable with parameter $p \in (0,1)$ if

$$\begin{aligned}P\{X=1\} &= p, \\ P\{X=0\} &= 1-p.\end{aligned}$$

Bernoulli random variables are quite useful because they are the building blocks for more complicated random variables. For a Bernoulli random variable with a parameter of p,

$$\mathbb{E}[X] = p \quad \text{and} \quad \mathsf{Var}(X) = p(1-p).$$

For any event $A \in \mathcal{F}$, we define the indicator function 1_A, or I_A, to be equal to one if A occurs, and zero otherwise. That is,

$$1_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{if } \omega \notin A \end{cases}$$

1_A is a Bernoulli random variable with parameter $P(A)$.

Binomial random variables: Consider n independent repeated trials of a Bernoulli random variable. Let X be the number of "successes" (i.e. 1's) in the n trials. The range of X is $\{0, 1, \ldots, n\}$ and it can be shown that the probability mass function is

$$P\{X = k\} = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & , \quad \text{if } k \in \{0, 1, 2, \ldots, n\} \\ 0 & , \quad \text{else} \end{cases} . \tag{2.2}$$

Any random variable with probability mass function (2.2) is a *binomial* random variable with parameters n and p.

Example 2.3.14. From the interval $(0, 1)$, 10 points are selected at random. What is the probability that at least 5 of them are less than 1/3?

Solution: A success is defined by $x_i < 1/3$ for $i \in \{1, 2, \ldots, 10\}$. Thus, $p = 1/3$. Let X be the number of successes. X is a binomial(10,1/3) random variable. The probability of at least 5 successes in 10 tries is then

$$P\{X \geq 5\} = \sum_{k=5}^{10} \binom{10}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{10-k} = 0.21312.$$

□

For a binomial random variable with parameters n and p,

$$\mathbb{E}[X] = np \quad \text{and} \quad \mathsf{Var}(X) = np(1-p).$$

Geometric random variables: Consider repeating a Bernoulli trial *until a success happens*. In this case, the sample space is

$$\Omega = \{s, fs, ffs, fffs, ...\},$$

where s denotes a success and f denotes a failure. Suppose that the probability of success is p. Let X be the number of trials until a success happens. The range of X is $\mathcal{R}(X) = \{1, 2, 3, ...\}$. The probability mass function is given by

$$P\{X = n\} = \begin{cases} (1-p)^{n-1} p, & n \in \{1, 2, 3, \ldots\} \\ 0 & \text{else} \end{cases}$$

Any random variable with this probability mass function is called a *geometric* random variable with a parameter of p. For a geometric random variable with a parameter of p,

$$\mathbb{E}[X] = \frac{1}{p} \quad \text{and} \quad \mathsf{Var}(X) = \frac{1-p}{p^2}.$$

Geometric random variables, along with exponential random variables, have the memoryless property. Let X be a Geometric(p) random variable. Then for all $n, m \geq 1$

$$P\{X > n+m \mid X > m\} = \frac{P\{X > n+m\}}{P\{X > m\}} = \frac{n+m \text{ failures to start}}{m \text{ failures to start}}$$
$$= \frac{(1-p)^{n+m}}{(1-p)^m} = (1-p)^n = P\{X > n\}.$$

In words, this says that the probability that the next n trials will be failures, given that the first m trials were failures, is the same as the probability that first n are failures.

Poisson random variables: This is one of the most important random variables in the study of stochastic models of reaction networks and will arise time and time again in this class.

A random variable with range $\{0, 1, 2, \dots\}$ is a *Poisson random variable* with parameter $\lambda > 0$ if

$$P\{X = k\} = \begin{cases} \frac{\lambda^k e^{-\lambda}}{k!} & , \quad k = 0, 1, 2, \dots \\ 0 & , \quad \text{else} \end{cases}.$$

For a Poisson random variable with a parameter of λ,

$$\mathbb{E}[X] = \lambda \quad \text{and} \quad \mathsf{Var}(X) = \lambda.$$

The Poisson random variable, together with the Poisson process, will play a central role in this class, especially when we discuss continuous time Markov chains. The following can be shown by, for example, generating functions.

Theorem 2.3.15. *If $X \sim$ Poisson(λ) and $Y \sim$ Poisson(μ) then $Z = X + Y \sim$ Poisson($\lambda + \mu$).*

The Poisson process. An increasing process on the integers $Y(t)$ is said to be a *Poisson process* with *intensity* (or *rate* or *propensity*) λ if

1. $Y(0) = 0$.
2. The number of events in disjoint time intervals are independent.
3. $P\{Y(s+t) - Y(t) = i\} = e^{-\lambda s}\frac{(\lambda s)^i}{i!}$, $i = 0, 1, 2, \dots$ for any $s, t \geq 0$

The Poisson process will be *the* main tool in the development of (pathwise) stochastic models of biochemical reaction systems in this class and will be derived more rigorously later.

2.3.4 Some common continuous random variables

Uniform random variables. Uniform random variables will play a central role in efficiently generating different types of random variables for simulation methods. Consider an interval (a, b), where we will often have $a = 0$ and $b = 1$. The random variable is said to be uniformly distributed over (a, b) if

$$F(t) = \begin{cases} 0 & t < 1 \\ (t-a)/(b-a) & a \le t < b \\ 1 & t \ge b \end{cases},$$

$$f(t) = F'(t) = \begin{cases} 1/(b-a) & a < t < b \\ 0 & else \end{cases}.$$

For a uniform random variable over the interval (a, b),

$$\mathbb{E}[X] = \frac{a+b}{2} \quad \text{and} \quad \mathsf{Var}(X) = \frac{(b-a)^2}{12}.$$

Normal random variables. Also called *Gaussian* random variables, normal random variables play a central role in the theory of probability due to their connection to the central limit theorem and Brownian motions. These connections will arise in this class when we consider diffusion approximations, called Langevin approximations in some of the sciences, to the continuous time Markov chain models of chemically reacting species.

A random variable X is called a *normal* with mean μ and variance σ^2, and we write $X \sim N(\mu, \sigma^2)$, if its density is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}.$$

A *standard normal* random variable is a normal random variable with $\mu = 0$ and $\sigma = 1$. For a normal random variable with parameters μ and σ^2,

$$\mathbb{E}[X] = \mu \quad \text{and} \quad \mathsf{Var}(X) = \sigma^2.$$

Exponential random variables. For reasons to be demonstrated later, the exponential random variable will be the most important continuous random variable in the study of continuous time Markov chains. It will turn out to be linked to the Poisson process in that the inter-event times of a Poisson process will be given by an exponential random variable. This will lead to the important fact that many simulation methods will consists of generating a sequence of correctly chosen exponential random variables.

A random variable X has an *exponential distribution* with parameter $\lambda > 0$ if it has a probability density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & , \quad x \ge 0 \\ 0 & , \quad \text{else} \end{cases}.$$

For an exponential random variable with a parameter of $\lambda > 0$,

$$\mathbb{E}[X] = \frac{1}{\lambda} \quad \text{and} \quad \mathsf{Var}(X) = \frac{1}{\lambda^2}.$$

Similar to the geometric random variable, the exponential random variable has the memoryless property.

Proposition 2.3.16 (Memoryless property)**.** *Let $X \sim Exp(\lambda)$, then for any $s, t \geq 0$,*

$$P\{X > (s+t) \mid X > t\} = P\{X > s\}. \tag{2.3}$$

Probably the most important role the exponential random variable will play in these notes is as the *inter-event time* of Poisson random variables.

Proposition 2.3.17. *Consider a Poisson process with rate $\lambda > 0$. Let T_i be the time between the ith and $i+1$st events. Then $T_i \sim Exp(\lambda)$.*

The following propositions are relatively straightforward to prove and form the heart of the usual methods used to simulate continuous time Markov chains. This method is often termed the *Gillespie algorithm* in the biochemical community, and will be discussed later in the notes.

Proposition 2.3.18. *If for $i = 1, \ldots, n$, the random variables $X_i \sim Exp(\lambda_i)$ are independent, then*

$$X_0 \equiv \min_i\{X_i\} \sim Exp(\lambda_0), \quad \textit{where} \quad \lambda_0 = \sum_{i=1}^{n} \lambda_i.$$

Proof. Let $X_0 = \min_i\{X_i\}$. Set $\lambda_0 = \sum_i \lambda_{i=1}^n$. Then,

$$P\{X_0 > t\} = P\{X_1 > t, \ldots, X_n > t\} = \prod_{i=1}^{n} P\{X_i > t\} = \prod_{i=1}^{n} e^{-\lambda_i t} = e^{-\lambda_0 t}.$$

□

Proposition 2.3.19. *For $i = 1, \ldots, n$, let the random variables $X_i \sim Exp(\lambda_i)$ be independent. Let j be the index of the smallest of the X_i. Then j is a discrete random variable with probability mass function*

$$P\{j = i\} = \frac{\lambda_i}{\lambda_0}, \quad \textit{where} \quad \lambda_0 = \sum_{i=1}^{n} \lambda_i.$$

Proof. We first consider the case of $n = 2$. Let $X \sim Exp(\lambda)$ and $Y \sim Exp(\mu)$ be independent. Then,

$$P\{X < Y\} = \iint_{x<y} \lambda e^{-\lambda x} \mu e^{-\mu y} dx dy = \int_0^\infty \int_0^y \lambda e^{-\lambda x} \mu e^{-\mu y} dx dy = \frac{\lambda}{\mu + \lambda}.$$

Now, returning to the general case of arbitrary n, we let $Y_i = \min_{j \neq i}\{S_j\}$, so that Y_i is exponential with rate $\sum_{j \neq i} \lambda_j$ by Proposition 2.3.18. Using the case $n = 2$ proved above then yields

$$P\{X_i < \min_{j \neq i}\{X_j\}\} = P\{X_i < Y_i\} = \frac{\lambda_i}{\lambda_i + \sum_{j \neq i} \lambda_j} = \frac{\lambda_i}{\sum_{j=1}^n \lambda_j}.$$

□

One interpretation of the above two propositions is the following. If you have n alarm clocks, with the ith set to go off after an $\text{Exp}(\lambda_i)$ amount of time, then Proposition 2.3.18 tells you when the first will go off, and Proposition 2.3.19 tells you which one will go off at that time.

2.3.5 Transformations of random variables

Most software packages have very good and efficient methods for the generation of pseudo-random numbers that are uniformly distributed on the interval $(0, 1)$. These pseudo random numbers are so good that we will take the perspective throughout these notes that they are, in fact, truly uniformly distributed over $(0, 1)$. We would then like to be able to construct all other random variables as transformations, or functions, of these uniform random variables. The method for doing so will depend upon whether or not the desired random variable is continuous or discrete. In the continuous case, Theorem 2.3.20 will often be used, whereas in the discrete case Theorem 2.3.22 will be used.

Theorem 2.3.20. *Let U be uniformly distributed on the interval $(0, 1)$ and let F be an invertible distribution function. Then $X = F^{-1}(U)$ has distribution function F.*

Before proving the theorem, we show how it may be used in practice.

Example 2.3.21. Suppose that we want to be able to generate an exponential random variable with parameter $\lambda > 0$. Such a random variable has distribution function $F : \mathbb{R}_{\geq 0} \to [0, 1)$

$$F(t) = 1 - e^{-\lambda t}, \quad t \geq 0.$$

Therefore, $F^{-1} : [0, 1) \to \mathbb{R}_{\geq 0}$ is given by

$$F^{-1}(u) = -\frac{1}{\lambda} \ln(1 - u), \quad 0 \leq u \leq 1.$$

If U is uniform$(0, 1)$, then so is $1 - U$. Thus, to simulate a realization of $X \sim \text{Exp}(\lambda)$, you first simulate U from uniform$(0, 1)$, and then set

$$x = -\frac{1}{\lambda} \ln(U) = \ln(1/U)/\lambda.$$

□

Proof. (of Theorem 2.3.20) Letting $X = F^{-1}(U)$ where U is uniform$(0,1)$, we have

$$\begin{aligned} P\{X \leq t\} &= P\{F^{-1}(U) \leq t\} \\ &= P\{U \leq F(t)\} \\ &= F(t). \end{aligned}$$

□

Theorem 2.3.22. *Let U be uniformly distributed on the interval $(0,1)$. Suppose that $p_k \geq 0$ for each $k \in \{0,1,\dots,\}$, and that $\sum_k p_k = 1$. Define*

$$q_k = P\{X \leq k\} = \sum_{i=0}^{k} p_i.$$

Let

$$X = \min\{k \mid q_k \geq U\}.$$

Then,

$$P\{X = k\} = p_k.$$

Proof. Taking $q_{-1} = 0$, we have for any $k \in \{0,1,\dots\}$,

$$P\{X = k\} = P\left\{q_{k-1} < U \leq q_k\right\} = q_k - q_{k-1} = p_k.$$

□

In practice, the above theorem is typically used by repeatedly checking whether or not $U \leq \sum_{i=0}^{k} p_i$, and stopping the first time the inequality holds. We note that the theorem is stated in the setting of an infinite state space, though the analogous theorem holds in the finite state space case.

2.3.6 More than one random variable

To discuss more than one random variable defined on the same probability space $(\Omega, \mathcal{F}, P)$, we need joint distributions.

Definition 2.3.23. Let $X_1, \dots, X_n$ be discrete random variables with domain Ω. Then

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

is called the *joint probability mass function* of $X_1, \dots, X_n$.

Definition 2.3.24. We say that $X_1, \dots, X_n$ are *jointly continuous* if there exists a function $f(x_1,\dots,x_n)$, defined for all reals, such that for all $A \subset \mathbb{R}^n$

$$P\{(X_1,\dots,X_n) \in A\} = \int \cdots \int_{(x_1,\dots,x_n \in A)} f(x_1,\dots,x_n) dx_1 \dots dx_n.$$

The function $f(x_1,\dots,x_n)$ is called the *joint probability density function.*

Expectations are found in the obvious way.

Theorem 2.3.25. *If $h : \mathbb{R}^n \to \mathbb{R}$ then*

$$\mathbb{E}[h(X_1, \ldots, X_n)] = \sum_{x_1 \in \mathcal{R}(X_1)} \cdots \sum_{x_n \in \mathcal{R}(X_n)} h(x_1, \ldots, x_n) p_{X_1,\ldots,X_N}(x_1, \ldots, x_n).$$

Corollary 2.3.26. *For random variables X and Y on the same probability space*

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

Definition 2.3.27. The random variables X and Y are **independent** if for any sets of real numbers A and B

$$P\{X \in A, Y \in B\} = P\{X \in A\} P\{Y \in B\}.$$

This implies that X and Y are independent if and only if

$$p(x, y) = p_X(x) p_Y(y)$$
$$f(x, y) = f_X(x) f_Y(y),$$

for discrete and continuous random variables, respectively.

Theorem 2.3.28. *Let X and Y be independent random variables and $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ be real valued functions; then $g(X)$ and $h(Y)$ are also independent random variables.*

Theorem 2.3.29. *Let X and Y be independent random variables. Then for all real valued functions g and h,*

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)].$$

One important application of the above theorem is the relation

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

if X and Y are independent. However, the converse is, in general, false.

Example 2.3.30. Let $R(X) = \{-1, 0, 1\}$ with $p(-1) = p(0) = p(1) = 1/3$. Let $Y = X^2$. We have

$$\mathbb{E}[X] = 0, \quad \mathbb{E}[Y] = 2/3, \quad \text{and} \quad \mathbb{E}[XY] = 0.$$

However,

$$P\{X = 1, Y = 1\} = P\{Y = 1 | X = 1\} P\{X = 1\} = 1/3$$
$$P\{X = 1\} P(Y = 1\} = (1/3) \times (2/3) = 2/9,$$

demonstrating these are *not* independent random variables. □

More generally, if $X_1, X_2, \ldots, X_n$ are independent random variables, then

$$\mathbb{E}[X_1 \cdots X_n] = \mathbb{E}[X_1] \cdots \mathbb{E}[X_n].$$

2.3.7 Variance of linear combinations.

Suppose that

$$X = X_1 + X_2 + \cdots + X_n.$$

We already know that for any X_i defined on the same probability space

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

For the variance of a linear combination, a direct calculation shows that for $a_i \in \mathbb{R}$,

$$\mathsf{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \mathsf{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{i<j} a_i a_j Cov(X_i, X_j),$$

where

$$Cov(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] = \mathbb{E}[X_i X_j] - 2\mu_i\mu_j + \mu_i\mu_j = \mathbb{E}[X_i X_j] - \mu_i\mu_j.$$

Therefore, if the X_i are pairwise independent,

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

Example 2.3.31. Let X be a binomial random variable with parameters n and p. Since X is the number of successes in n independent trials, we can write

$$X = X_1 + \cdots + X_n,$$

where X_i is 1 if ith trial was success, and zero otherwise. Therefore, the X_i's are independent Bernoulli random variables and $\mathbb{E}[X_i] = P\{X_i = 1\} = p$. Thus,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p = np.$$

Because each of the X_i's are independent

$$\mathsf{Var}(X) = \sum_{i=1}^{n} \mathsf{Var}(X_i) = \sum_{i=1}^{n} p(1-p) = np(1-p).$$

□

Proposition 2.3.32. *Let $X_1, \ldots, X_n$ be n independent random variables with mean μ and variance σ^2. Let $\bar{X} = (1/n)(X_1 + \cdots X_n)$ be the average of the sample. Then*

$$\mathbb{E}(\bar{X}) = \mu, \qquad \mathsf{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Proof. Calculating shows

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n} n\mu = \mu$$

$$\mathsf{Var}(\bar{X}) = \mathsf{Var}\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

□

2.4 Inequalities and Limit Theorems

2.4.1 Important inequalities

Oftentimes we do not have explicit representations for the distributions, probability mass functions, or densities of the random variables of interest. Instead, we may have means, variances, or some other information. We may use this information to garner some bounds on probabilities of events.

Theorem 2.4.1 (Markov's Inequality)**.** *Let X be a non-negative random variable; then for any $t > 0$*

$$P\{X \geq t\} \leq \frac{\mathbb{E}[X]}{t}.$$

Proof. The proof essentially uses the indicator functions $1_{\{X \geq t\}}$ and $1_{\{X < t\}}$ to break up X into two pieces:

$$\mathbb{E}[X] = \mathbb{E}[X1_{\{X \geq t\}}] + \mathbb{E}[X(1 - 1_{\{X \geq t\}})] \geq \mathbb{E}[X1_{\{X \geq t\}}] \geq \mathbb{E}[t1_{\{X \geq t\}}] = tP\{X \geq t\}.$$

□

Theorem 2.4.2 (Chebyshev's Inequality)**.** *If X is a random variable with mean μ and variance $\sigma^2 < \infty$, then for any $t > 0$,*

$$P\{|X - \mu| \geq t\} \leq \frac{\sigma^2}{t^2}.$$

Proof. We have that $(X - \mu)^2 \geq 0$, so we may use the Markov inequality:

$$P\left\{(X - \mu)^2 \geq t^2\right\} \leq \frac{E\left[(X - \mu)^2\right]}{t^2} = \frac{\sigma^2}{t^2}.$$

But, $P\left\{(X - \mu)^2 \geq t^2\right\} = P\left\{|X - \mu| \geq t\right\}$, and so the result is shown. □

2.4.2 Limit theorems

We now present three limit theorems that will be used later in the course.

Theorem 2.4.3 (Weak Law of Large Numbers)**.** *Let $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed random variables with $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \mathsf{Var}(X_i) < \infty$, $i = 1, 2, \ldots$. Then for all $\epsilon > 0$*

$$\lim_{n \to \infty} P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right\} = 0.$$

Proof. Let $\overline{X} = (1/n)\sum_i X_i$ be the sample average. We know that $\mathbb{E}\overline{X} = \mu$ and $\mathsf{Var}(\overline{X}) = \sigma^2/n$. Thus, by Chebyshev's inequality we get

$$P\left\{\left|\overline{X} - \mu\right| > \epsilon\right\} \leq \frac{\sigma^2/n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ as } n \to \infty.$$

□

Note that the proof gives a rate of convergence of $O(1/n)$. The corresponding *strong law of large numbers* is now stated.

Theorem 2.4.4 (Strong law of large numbers)**.** *Let $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed random variables with mean μ. then*

$$P\left\{\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right\} = 1.$$

So $\bar{X} = (X_1 + \cdots X_n)/n$ converges to μ *almost surely*, or with a probability of one.

We now state the central limit theorem.

Theorem 2.4.5 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each with expectation μ and variance σ^2. Then the distribution of*

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

converges to the distribution of a standard normal random variable. That is, for any $t \in (-\infty, \infty)$

$$\begin{aligned}\lim_{n\to\infty} P\{Z_n \le t\} &= \lim_{n\to\infty} P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le t\right\} \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx.\end{aligned}$$

2.5 Simulation

Consider the following question: given a random variable X with unknown distribution function $F(x)$, how can we estimate $\mu = \mathbb{E}[X]$?

Assuming that we can generate realizations of X via a computer, the simulation approach to solving this problem is to estimate $\mu = \mathbb{E}[X]$ by running n independent and identical experiments, thereby obtaining n i.i.d. random variables $X_1, X_2, \ldots, X_n$, with each having the distribution $F(x)$. Then, take the estimate as

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

We call μ_n an *estimator*. In this case, we have an *unbiased estimator* as

$$\mathbb{E}\mu_n = \frac{1}{n}\mathbb{E} \sum_{i=1}^{n} X_i = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \mu.$$

Further, by the strong law of large numbers we know that

$$\mu_n \to \mu,$$

as $n \to \infty$, with a probability of one.

Of course, knowing that $\mu_n \to \mu$, as $n \to \infty$, does not actually tell us how large of an n we need in practice. This brings us to the next logical question: how good is the estimate for a given, finite n. To answer this question, we will apply the central limit theorem.

We know from the central limit theorem that

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \overset{D}{\approx} N(0,1),$$

or

$$\frac{\sqrt{n}}{\sigma}(\mu_n - \mu) \overset{D}{\approx} N(0,1).$$

Specifically, for any $z \in \mathbb{R}$

$$\begin{aligned} P\left\{-z \le N(0,1) \le z\right\} &\approx P\left\{-z \le \frac{\sqrt{n}}{\sigma}(\mu_n - \mu) \le z\right\} \\ &= P\left\{-\frac{\sigma z}{\sqrt{n}} \le (\mu_n - \mu) \le \frac{\sigma z}{\sqrt{n}}\right\} \\ &= P\left\{\mu_n - \frac{\sigma z}{\sqrt{n}} \le \mu \le \mu_n + \frac{\sigma z}{\sqrt{n}}\right\}. \end{aligned}$$

In words, the above says that the probability that the true value, μ, is within $\pm\sigma z/\sqrt{n}$ of the estimator μ_n is $P\left\{-z \le N(0,1) \le z\right\}$, which can be found for any z. More importantly, a value z can be found for any desired level of confidence. The interval $(\mu - \sigma z/\sqrt{n}, \mu + \sigma z/\sqrt{n})$ is called our *confidence interval* and the probability $P\left\{-z \le N(0,1) \le z\right\}$ is our *confidence*. Note that both our confidence and the size of the confidence interval increase as z is increased.

We now turn to finding the value z for a desired confidence level. Letting

$$\Phi(z) = P\{N(0,1) \le z\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt,$$

we have

$$\begin{aligned} P\{-z \le N(0,1) \le z\} &= P\{N(0,1) \le z\} - P\{N(0,1) \le -z\} \\ &= \Phi(z) - \Phi(-z) \\ &= \Phi(z) - (1 - \Phi(z)) \\ &= 2\Phi(z) - 1. \end{aligned}$$

Therefore, if for some $\delta > 0$ we want to have a probability of $1 - \delta$ that the true value is in the constructed confidence interval, then we must choose z so that

$$P\{-z \le N(0,1) \le z\} = 1 - \delta.$$

That is, we need to find a z so that

$$2\Phi(z) - 1 = 1 - \delta,$$

or

$$\Phi(z) = 1 - \frac{\delta}{2}.$$

For example, if $\delta = .1$, so that a 90% confidence interval is required, then

$$\Phi(z) = 1 - .05 = .95,$$

and $z = 1.65$. If, on the other hand, we want $\delta = 0.05$, so that a 95% confidence interval is desired, then

$$\Phi(z) = 1 - .025 = .975$$

and $z = 1.96$.

Summarizing, we see that for a given δ, we can find a z so that the probability that the parameter μ, which is what we are after, lies in the interval

$$\left[\mu_n - \frac{\sigma z}{\sqrt{n}},\ \mu_n + \frac{\sigma z}{\sqrt{n}}\right]$$

is approximately $1-\delta$. Further, as n gets larger, the confidence interval shrinks. It is worth pointing out that the confidence interval shrinks at a rate of $1/\sqrt{n}$. Therefore, to get a 10-fold increase accuracy, we need a 100-fold increase in work.

There is a major problem with the preceding arguments: if we don't know μ, which is what we are after, we most likely do not know σ either. Therefore, we will also need to estimate it from our independent samples $X_1, X_2, \ldots, X_n$.

Theorem 2.5.1. *Let $X_1, \ldots, X_n$ be independent and identical samples with mean μ and variance σ^2, and let*

$$\sigma_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_n)^2,$$

where μ_n is the sample mean

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then,

$$\mathbb{E}\sigma_n^2 = \sigma^2.$$

Proof. We have

$$\mathbb{E}\sigma_n^2 = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_n)^2\right],$$

which yields

$$\begin{aligned}(n-1)\mathbb{E}\sigma_n^2 = \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu_n)^2\right] &= \sum_{i=1}^{n}\mathbb{E}[X_i^2] - 2\mathbb{E}\left[\mu_n \sum_{i=1}^{n} X_i\right] + n\mathbb{E}\left[\mu_n^2\right] \\ &= \sum_{i=1}^{n}\mathbb{E}[X_i^2] - 2\mathbb{E}\left[\mu_n n \mu_n\right] + n\mathbb{E}\left[\mu_n^2\right] \\ &= n\mathbb{E}[X^2] - n\mathbb{E}\left[\mu_n^2\right].\end{aligned}$$

However,

$$\mathbb{E}\left[\mu_n^2\right] = \mathsf{Var}(\mu_n) + \mathbb{E}\left[\mu_n\right]^2 = \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + \mu^2 = \frac{1}{n}\sigma^2 + \mu^2.$$

Therefore,

$$\frac{n-1}{n}\mathbb{E}\sigma_n^2 = \mathbb{E}[X^2] - \mathbb{E}[\mu_n^2] = \left(\sigma^2 + \mu^2\right) - \left(\frac{1}{n}\sigma^2 + \mu\right) = \frac{n-1}{n}\sigma^2,$$

completing the proof. □

Therefore, we can use

$$\sigma_n = \sqrt{\sigma_n^2}$$

as an estimate of the standard deviation in the confidence interval and

$$\left[\mu_n - \frac{\sigma_n z}{\sqrt{n}}, \quad \mu_n + \frac{\sigma_n z}{\sqrt{n}}\right]$$

is an approximate $(1-\delta)100\%$ confidence interval for $\mu = \mathbb{E}[X]$.

We note that there are two sources of error in the development of the above confidence interval that we will not explore here. First, there is the question of how good an approximation the central limit theorem is giving us. For reasonably sized n, this should not give too much of an error. The second source of error is in using σ_n as opposed to σ. Again, for large n, this error will be relatively small.

We have the following algorithm for producing a confidence interval for an expectation given a number of realizations.

Algorithm for producing confidence intervals for a given n.

1. Select n, the number of experiments to be run, and $\delta > 0$.

2. Perform n independent replications of the experiment, obtaining the observations $X_1, X_2, \ldots, X_n$ of the random variable X.

3. Compute the sample mean and sample variance

$$\mu_n = \frac{1}{n}(X_1 + \cdots + X_n)$$
$$\sigma_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_n)^2.$$

4. Select z such that $\Phi(z) = 1-\delta/2$. Then an approximate $(1-\delta)100\%$ confidence interval for $\mu = \mathbb{E}[X]$ is

$$\left[\mu_n - \frac{\sigma_n z}{\sqrt{n}}, \quad \mu_n + \frac{\sigma_n z}{\sqrt{n}}\right].$$

If a level of precision is desired, and n is allowed to depend upon δ and a tolerance ϵ, then the following algorithm is most useful.

Algorithm for producing confidence intervals to a given tolerance.

1. Select $\delta > 0$, determining the desired confidence, and $\epsilon > 0$ giving the desired precision. Select z such that $\Phi(z) = 1 - \delta/2$.

2. Perform independent replications of the experiment, obtaining the observations $X_1, X_2, \ldots, X_n$ of the random variable X, until
$$\frac{\sigma_n z}{\sqrt{n}} < \epsilon.$$

3. Report
$$\mu_n = \frac{1}{n}(X_1 + \cdots + X_n)$$
and the $(1-\delta)100\%$ confidence interval for $\mu = \mathbb{E}[X]$,
$$\left[\mu_n - \frac{\sigma_n z}{\sqrt{n}}, \quad \mu_n + \frac{\sigma_n z}{\sqrt{n}}\right] \approx [\mu - \epsilon, \mu + \epsilon].$$

There are two conditions usually added to the above algorithm. First, there is normally some minimal number of samples generated, n_0 say, before one checks whether or not $\sigma_n z/\sqrt{n} < \epsilon$. Second, it can be time consuming to compute the standard deviation after every generation of a new iterate. Therefore, one normally only does so after every multiple of M iterates, for some positive integer $M > 0$. We will not explore the question of what "good" values of n_0 and M are, though taking $n_0 = M \approx 100$ is usually sufficient.

2.6 Exercises

1. Verify, through a direct calculation, Equation (2.1).

2. Verify the memoryless property, Equation (2.3), for exponential random variables.

3. Matlab exercise. Perform the following tasks using Matlab.

   (a) Using a FOR LOOP, use the etime command to time how long it takes Matlab to generate 100,000 exponential random variables with a parameter of 1/10 using the built-in exponential random number generator. Sample code for this procedure is provided on the course website.

   (b) Again using a FOR LOOP, use the etime command to time how long it takes Matlab to generate 100,000 exponential random variables with parameter 1/10 using the transformation method given in Theorem 2.3.20.

4. Matlab exercise. Let X be a random variable with range $\{-10,\ 0,\ 1,\ 4,\ 12\}$ and probability mass function

$$P\{X=-10\}=\frac{1}{5},\quad P\{X=0\}=\frac{1}{8},\quad P\{X=1\}=\frac{1}{4},$$
$$P\{X=4\}=\frac{1}{3},\quad P\{X=12\}=\frac{11}{120}.$$

Using Theorem 2.3.22, generate N independent copies of X and use them to estimate $\mathbb{E}X$ via

$$\mathbb{E}X \approx \frac{1}{N}\sum_{i=1}^{N} X_{[i]},$$

where $X_{[i]}$ is the ith independent copy of X and $N \in \{100, 10^3, 10^4, 10^5\}$. Compare the result for each N to the actual expected value. A helpful sample Matlab code has been provided on the course website.