PROBABILITY COURSE NOTES

STAT/MATH 394/395

Nathaniel Derby

nderby@stat.washington.edu

Summer 2009

University of Washington

Table of Contents

Notation

The following notation will be used consistently throughout these notes:

$\mathbb{Z}$	The set of integers: $\{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$.
$\mathbb{Z}^+$	The set of natural, or counting, numbers: $\{1, 2, 3, \ldots\}$.
$\mathbb{Z}^*$	The set of whole numbers: $\{0, 1, 2, 3, \ldots\}$.
$\mathbb{Z}^-$	The set of negative integers: $\{\ldots, -3, -2, -1\}$.
$\mathbb{R}$	The set of real numbers.
$\mathbb{R}^+$	The set of positive real numbers.
$\mathbb{R}^-$	The set of negative real numbers.
$\mathbb{Q}$	The set of rational numbers: $\left\{\frac{a}{b} : \ a, b \in \mathbb{Z}\right\}$.
$\varnothing$	The empty set (set with no elements in it).
$\cup$	Union (of two sets).
$\cap$	Intersection (of two sets).
$\forall$	"For all"
$\exists$	"There exists"
$\ni$	"such that"

In some other texts, $\mathbb{N}$ is used to denote the set of natural numbers $\mathbb{Z}^+$. This notation is not used in this text, but it is valid and commonly used.

1 Algebra of Sets

1.1 Elementary Set Theory

Definition 1.1.1. A *set* is a collection of objects.

Examples of sets include

- The set of all positive integers: $\mathbb{Z}^+ = \{1, 2, 3, \ldots\}$.
- The set of all white mice in a room.

Sets are usually denoted by capital italic letters, such as A, B, N, I, S, etc.

Definition 1.1.2. An *element* or *member* of a set is an object that belongs to the set in question. If b is an element of the set S, then we use the notation $b \in S$.

Definition 1.1.3. A set is *completely described* if a rule is given that determines whether any particular object is an element of that set.

Example 1.1.1. Let A be the set of numbers 2, 4 and 6. Then we write $A = \{2, 4, 6\}$. This set is *finite*, since it has a finite number of elements. Note that $2 \in A$, $4 \in A$, $6 \in A$ but $8 \notin A$. The notation $\notin$ means "does not belong to". □

Definition 1.1.4. The *size* or *cardinality* of a finite set is the number of elements in the set. We use the notation $|A|$ to denote the size (or cardinality) of the set A.

Example 1.1.2. Let B be the set of all positive even integers:

$$B = \{2, 4, 6, 8, 10, \ldots\}.$$

This is not as clear as

$$B = \{x : \ x = 2n, \ n \text{ is a positive integer } \}$$

or

$$B = \{x : \ x = 2n, \ n \in \mathbb{Z}^+\}, \text{ where } \mathbb{Z}^+ \text{ denotes the set of positive integers.}$$

The set B is *infinite*, since it has infinitely many members. In fact, B is *countably infinite*. □

Definition 1.1.5. A *countably* or *denumerably infinite* set is an infinite set whose members can be labeled with subscripts such that every positive integer is used as a subscript exactly once. An infinite set which is not countably infinite is an *uncountably infinite* set.

Example 1.1.3. For example,

$$B = \{x : \ x = 2n, \ n \in \mathbb{Z}^+\}$$
$$Q = \{n : \ n \in \mathbb{Z}^-\}$$

are countably infinite sets. Examples of uncountably infinite sets are

$$C = \{x : \ 0 < x < 6, \ x \in \mathbb{R}\}$$
$$D = \{(x, y) : \ 0 \leq x \leq 1, \ 0 \leq y \leq 1, \ x, y \in \mathbb{R}\}.$$

The set

$$E = \{(x, y) : \ 0 \leq x \leq 1, \ 0 \leq y \leq 1, \ x, y \in \mathbb{Z}^+\}$$

is a finite set. In fact, $E = \{(1, 1)\}$. □

Example 1.1.4. The members of a set need not have any recognizable properties in common. For example, consider the set

$$\{\text{University of Washington, Toronto, } 9\}.$$

Definition 1.1.6. If every member of a set A is also a member of a set B, then A is called a *subset* of B. Equivalently, A is *contained* in B. We use the notation $A \subset B$, or $A \subseteq B$, or analogously, $B \supset A$, or $B \supseteq A$.

Example 1.1.5. If $A = \{2, 4, 6\}$ and $B = \{2, 4, 6, 8, 10\}$, then $A \subset B$. Note two different concepts (and notations) here:

$\{2\} \subset A \quad \Rightarrow \quad$ The *set* which contains 2 as its only element is a *subset* of the set A
$2 \ \in A \quad \Rightarrow \quad$ The *element* 2 is a *member* of the set A □

Definition 1.1.7. The sets A and B are *equal*, denoted $A = B$, if and only if $A \subset B$ and $B \subset A$.

Definition 1.1.8. The *empty* or *null set* is the set which has no members. The empty set is denoted by either $\{\ \}$ (literally, a set with no members in it) or by $\varnothing$ (which is sometimes also written as $\emptyset$). By convention, the empty set is considered to be a subset of all sets. Thus, if A is any set, $\varnothing \subset A$.

Example 1.1.6. Let $S = \{a, b, c\}$. Then then subsets of S are

$$\varnothing, \ \{a\}, \ \{b\}, \ \{c\}, \ \{a, b\}, \ \{a, c\}, \ \{b, c\}, \ \{a, b, c\}. \tag{1.1.1}$$

□

If S is a finite set of size n (i.e., S has n members), then it can be shown that the number of possible subsets of S is 2^n. For this reason, the collection of all possible subsets of S is denoted as 2^S. For example, for S in Example 1.1.6, the collection of sets 2^S is given by (1.1.1). Note that it has $2^n = 2^3 = 8$ subsets.

Definition 1.1.9. A is a *proper subset* of B if $A \subset B$ but $A \neq B$ and $A \neq \varnothing$.

Definition 1.1.10. The *universal set*, denoted U, is a set which contains as subsets all sets within a specific discussion.

Example 1.1.7. If we want to discuss the number of live polio vaccines within a given batch of vaccines, we might let U be the set of all positive real numbers. That is, $U = \mathbb{Z}^+$. It would suffice, however, to let $U = \{x : x \in \mathbb{Z}^+, x \leq 10^{23}\}$. □

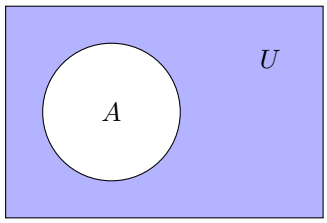
Definition 1.1.11. Let $A \subseteq U$. The set which consists of all elements of U which are not in A is called the *complement* of A with respect to (wrt) U, and is denoted by a few different notations, such as

$$\bar{A},\ A^c,\ A^*,\ \sim A,\ U-A,\ \text{or } U/A.$$

In these notes, we shall use the notation A^c.

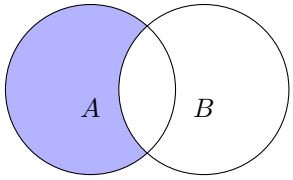
Note that for arbitrary sets A and B, B/A is used to denote the elements of B which are not in A. The set A need not be a subset of B.

Example 1.1.8 (Venn Diagram). Draw a geometric shape on the plane and assume that points interior to and on the boundary of the figure constitute the set.



$A^c = U/A$ is the shaded region.

□



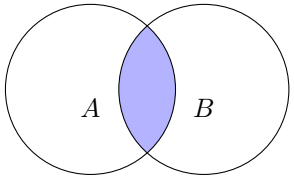
A/B is the shaded region.

Example 1.1.9. If $U = \{1,2,3\}$ and $A = \{1,2\}$, then $A^c = \{3\}$. If $B = \{2,3,4\}$, then $B/A = \{3,4\}$ and $A/B = \{1\}$. □

Definition 1.1.12. If A and B are sets, then A *intersection* B is the set of all elements which belong to both A and B. This is denoted by $A \cap B$, or simply AB.

Note that $A \cap B = B \cap A$. That is, order is not important; the operation is commutative.

Example 1.1.10. If $A = \{a,b,c,d\}$ and $B = \{a,d,f\}$, then $A \cap B = \{a,d\}$.

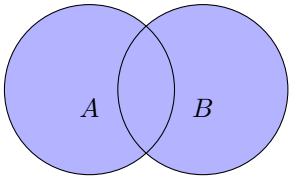


AB is the shaded region. □

Definition 1.1.13. If A and B are sets, then A *union* B is the set of all elements which belong to either A or B, or both. This is denoted by $A \cup B$.

Note that $A \cup B = B \cup A$. Again, the operation is commutative.

Example 1.1.11. If the sets $A = \{a, b, c, d\}$ and $B = \{a, d, f\}$, then $A \cup B = \{a, b, c, d, f\}$.



$A \cup B$ is the shaded region. □

Definition 1.1.14. Two sets are *disjoint* if (and only if) they have no elements in common. That is, if $A \cap B = \varnothing$.

Theorem 1.1.1 (DeMorgan's Laws).

$$\left(\bigcup_{i=1}^{n} E_i\right)^c = \bigcap_{i=1}^{n} E_i^c \tag{1.1.2}$$

$$\left(\bigcap_{i=1}^{n} E_i\right)^c = \bigcup_{i=1}^{n} E_i^c \tag{1.1.3}$$

Proof: Suppose $x \in (\cup_i E_i)^c$. Then $x \notin \cup_i E_i$, which means that $x \notin E_i \ \forall i$. Thus, $x \in E_i^c \ \forall i \quad \Rightarrow \quad x \in \cap_i E_i^c$. Now suppose that $x \in \cap_i E_i^c$. Then $x \in E_i^c \ \forall i \quad \Rightarrow \quad x \notin E_i \ \forall i \quad \Rightarrow \quad x \notin \cup_i E_i \quad \Rightarrow \quad x \in (\cup_i E_i)^c$. This proves (1.1.2).

Using (1.1.2),

$$\left(\bigcup_{i=1}^{n} E_i^c\right)^c = \bigcap_{i=1}^{n} (E_i^c)^c = \bigcap_{i=1}^{n} E_i$$

since $(E^c)^c = E$ for any set E. Taking the complement of both sides gives us (1.1.3). ■

Definition 1.1.15. An *n-tuple* $(x_1, x_2, \ldots, x_n)$ is an array of n symbols $x_1, x_2, \ldots, x_n$ which are called, respectively, the first component, the second component, ..., the n^{th} component.

Two n-tuples are equal iff they have the same symbols in the same order. That is,

$$(2, 7, 8) = (2, 7, 8),$$
$$(2, 7, 8) \neq (2, 8, 7).$$

However, in a set, the order of the elements is not important. That is,

$$\{2, 7, 8\} = \{2, 8, 7\}.$$

Note that the 3-tuple $(1, 0, 1)$ cannot be condensed, but the set $\{1, 0, 1\}$ is equal to $\{0, 1\}$ or $\{1, 0\}$.

1.2 Set Functions

In calculus, functions like

$$f(x) = 2x + 6, \qquad -\infty < x < \infty$$

or

$$h(x, y) = e^{-(x+y)}, \qquad x > 0,\ y > 0$$

are called *point functions*. They are evaluated at a point, such as x or (x, y). When we evaluate a function at a set instead of at a point, we call it a *set function*.

Example 1.2.1. Let A be a set of points in one-dimensional space (on the real line). Let $Q(A)$ be equal to the number of points in A which correspond to positive integers. $Q(\cdot)$ is a function of the set A. As examples,

$$\begin{aligned} A = \{x :\ 0 < x < 5\} \quad &\Rightarrow \quad Q(A) = 4, \\ A = \{x :\ x = -2, -1\} \quad &\Rightarrow \quad Q(A) = 0, \\ A = \{x :\ -\infty < x < 6\} \quad &\Rightarrow \quad Q(A) = 5. \end{aligned}$$

The domain of $Q(\cdot)$ is the set of subsets of the reals $\mathbb{R}$. The range of $Q(\cdot)$ is the set of non-negative integers $\mathbb{Z}^*$. □

Example 1.2.2. Let A be a set in two-dimensional space. Let $Q(A)$ be equal to the area of A if A has finite area, and undefined otherwise. As examples,

$$\begin{aligned} A = \{(x, y) :\ x^2 + y^2 < 1\} \quad &\Rightarrow \quad Q(A) = \pi, \\ A = \{(x, y) :\ (x, y) = (0, 0), (1, 1), (0, 1), (1, 0)\} \quad &\Rightarrow \quad Q(A) = 0, \\ A = \{(x, y) :\ x \geq 0,\ y \geq 0\} \quad &\Rightarrow \quad Q(A) \text{ is undefined.} \end{aligned}$$

□

Example 1.2.3. Let A be a set in one-dimensional space, and $Q(A) = \int_A e^{-x}\, dx$.

$$\begin{aligned} A = \{x :\ 0 < x < \infty\} \quad &\Rightarrow \quad Q(A) = \int_0^\infty e^{-x}\, dx = 1, \\ A = \{x :\ 0 < x < 1\} \quad &\Rightarrow \quad Q(A) = \int_0^1 e^{-x}\, dx = 1 - e^{-1}. \end{aligned}$$

□

Definition 1.2.1. A *class* is a set whose elements are all sets. The set of all subsets of a set is a class. We use script letters to denote classes; $\mathcal{A}$, $\mathcal{X}$, etc.

Definition 1.2.2. Let $\mathcal{Y}$ be a class of sets. A rule $\mu(\cdot)$ which associates with each $A \in \mathcal{Y}$ one and only one real number $\mu(A)$ is called a *real-valued set function* defined on $\mathcal{Y}$. The domain and range of $\mu(\cdot)$ are $\mathcal{Y}$ and $\mathbb{R}$, respectively.

Definition 1.2.3. A class $\mathcal{Y}$ is *closed* wrt some (binary) operation f iff for every $A, B \in \mathcal{Y}$, $AfB \in \mathcal{Y}$.

Example 1.2.4. Let $\mathcal{Y} = \{\varnothing, \{a\}\}$ and $f =$ union. Then $\mathcal{Y}$ is closed wrt f, since

$$\begin{aligned} \varnothing \cup \{a\} &= \{a\} \in \mathcal{Y}, \\ \{a\} \cup \{a\} &= \{a\} \in \mathcal{Y}, \\ \varnothing \cup \varnothing &= \varnothing \in \mathcal{Y}. \end{aligned}$$

If $f =$ intersection, then $\mathcal{Y}$ is closed wrt f, since

$$\begin{aligned} \varnothing \cap \{a\} &= \varnothing \in \mathcal{Y}, \\ \{a\} \cap \{a\} &= \{a\} \in \mathcal{Y}, \\ \varnothing \cap \varnothing &= \varnothing \in \mathcal{Y}. \end{aligned}$$

If $\mathcal{Y} = \{\{a\}, \{b\}\}$, then $\mathcal{Y}$ is not closed wrt $\cup$ nor wrt $\cap$. □

Is the set of all subsets of a finite set S closed wrt $\cup$ and $\cap$? We often require that $\mathcal{Y}$ be closed wrt one or more operations. This will be true when $\mathcal{Y}$ is the domain for a probability function.

2 Fundamental Definitions and Axioms

2.1 Random Experiments and Sample Spaces

Definition 2.1.1. A *random experiment* is one which can be repeated under the same conditions but whose outcome cannot be predicted with certainty.

Example 2.1.1. Examples of random experiments include experiments to determine the effect of a drug on a given type of patient experiments to determine the effect of a fertilizer on a given type of corn, and outcomes from games of poker. □

Definition 2.1.2. A *sample space* is the set of all possible individual outcomes to a given experiment (or game). We denote the sample space by S.

Example 2.1.2. If the game is tossing a coin in which the possible outcomes on each coin are a head (H) or a tail (T), then $S = \{H, T\}$. If a possibility is that the coin lands on its edge (E), then $S = \{H, T, E\}$. If the game is tossing two coin sin which the possible outcomes on each coin are H and T, then

$$\begin{aligned} &S = \{(H,H),(H,T),(T,H),(T;T)\}, \text{ or} \\ &S = \{(x,y) : x = H,T \text{ and } y = H,T\}, \text{ or} \\ &S = \{\{H,H\},\ \{H,T\},\ \{T,T\}\}. \end{aligned}$$

Note that curly brackets $\{\cdot,\cdot\}$ denote an unordered pair, while parentheses $(\cdot,\cdot)$ denote an ordered pair. Whether or not the points in the sample space are ordered depends on the questions asked about possible outcomes. If possible outcomes are ordered, the sample space must be ordered. □

Example 2.1.3. A sample space need not be finite. Consider the following:

(a) A Geiger counter set up to record cosmic ray counts. The number of counts recorded may be any positive integer or zero:
$$S = \{0, 1, 2, 3, \ldots\} = \mathbb{Z}^+ \cup \{0\} = \mathbb{Z}^*.$$

(b) An experiment measuring the time (in nanoseconds) between two neighboring peaks on an electrocardiogram:
$$S = \{x : 0 < x < \infty\} = \mathbb{R}^+.$$

Note that *S is not unique.* A sample space must contain all possible points, but it may also contain points or intervals which could never occur. One may make a sample space as large as one pleases at the price of having a large number of points (or intervals) in S which have zero probability associated with them.

Definition 2.1.3 (Population). A *population* is a set of objects from which we sample. We measure some (or many) properties of the object drawn from the population.

If the random experiment is tossing a coin, then the population must be the set of all coins, the set of all coins of the type being tossed, the set of all coins in a given box from which the tossed coin was drawn, or the set consisting of the coin being tossed.

Example 2.1.4. If the experiment is to determine the effect of anesthetics on patients, then the population might be the set of all patients who have had one of the different types of this anesthetic. When a patient is drawn from the population, we might measure the anesthetic type given, the age of the patient, and whether the operation was a success or failure. The sample space is the set of elements of the type (x, y, z), where

x denotes the type of anesthetic given to a certain patient,
y denotes the age of the patient, and
z denotes success or failure.

If three types (a, b, c) of anesthetic are used, then the sample space S might be

$$S = \{(x, y, z) : x = a, b, c;\ y = 1, \ldots, 100;\ z = s, f\}.$$

□

Example 2.1.5. If we want to measure the sugar content of oranges from a given grove, we could let the population be a particular harvest from the grove. A number of oranges would be measured for sugar content. The sample space in this case could be

$$S = \{x :\ 0 \leq x < \infty\}.$$

□

2.2 Events

Definition 2.2.1. An *event* is a subset of the sample space.

If upon performance of a random experiment the observed outcome is contained in some subset C of the sample space S, we say that *the event C has occurred.*

Example 2.2.1. If $S = \{(H, H), (T, T), (H, T), (T, H)\}$, then the $2^4 = 16$ possible events are the following:

$$\begin{aligned} A_1 &= S \\ A_2 &= \varnothing \\ A_3 &= \{(H, H)\} \\ A_4 &= \{(T, T)\} \\ A_5 &= \{(H, H), (T, T)\} \\ &\vdots \end{aligned}$$

If a (H, H) occurs, the events A_1, A_3, A_5 among others have occurred. □

Example 2.2.2. In Example 2.1.3(b), $S = \{x :\ 0 < x < \infty\} = \mathbb{R}^+$. One might be interested in the events

$$\begin{aligned} A &= \{x :\ x > 3\}, \text{ or} \\ B &= \{x :\ 1 < x < 4\}. \end{aligned}$$

□

Example 2.2.3. In Example 2.1.4, $S = \{(x, y, z) : x = a, b, c;\ y = 1, \ldots, 100;\ z = s, f\}$. One might be interested in the events

$$A = \{(x, y, z) :\ x = a;\ y = 1, \ldots, 100;\ z = s\},\ \text{or}$$
$$B = \{(x, y, z) :\ x = a;\ y = 1, \ldots, 50;\ z = f\}.$$ □

2.3 Probability Measure

Let S be the sample space for a random experiment. Let $C \subset S$ be an event. Assume that we have made n repeated performances (trials) of a random experiment. We count the number f of times that the event C actually occurred throughout the n trials. The ratio f/n is called the *relative frequency* of the event C in the n experiments.

Relative frequency is usually quite erratic for small values of n. As n increases, experience indicates that the relative frequency tends to stabilize. That is,

$$\lim_{n\to\infty} \frac{f}{n} = p.$$

We associate with the event C the number p. Thus, although we cannot predict the outcome of a random experiment, we can, for a large value of n, predict (approximately) the relative frequency with which the outcome will be in C.

Example 2.3.1. The number of white balls drawn in 600 trials of the experiment of drawing a ball with replacement from an urn containing 4 white and 2 red balls:

In trials numbered	Number of white balls drawn	Probability of white balls
001 - 100	69	$\frac{69}{100} = 0.690$
101 - 200	70	$\frac{69+70}{200} = 0.695$
201 - 300	59	$\frac{69+70+59}{300} = 0.660$
301 - 400	63	$\frac{69+70+59+63}{400} = 0.652$
401 - 500	76	$\frac{69+70+59+63+76}{500} = 0.674$
501 - 600	64	$\frac{69+70+59+63+76+64}{600} = 0.668$

□

Let S denote the sample space. It is our purpose to define a set function $\mathrm{P}(C)$ so that if $C \subset S$ is a probabilizable event, then $\mathrm{P}(C)$ is the probability that the outcome of the random experiment is an element of C.

We need to keep the structure of each set C sufficiently simple to allow computation of $\mathrm{P}(C)$. We ensure this by our definition of the domain of the set function $P(\cdot)$.

Let Ψ denote the domain of the set function $\mathrm{P}(\cdot)$. If S is finite or countably infinite, then Ψ is the set of all subsets of S. If S is uncountably infinite, then we require that Ψ is a non-empty set of subsets of S which is closed wrt countable unions and complements (i.e., if $A \in \Psi$, then $A^c = S/A \in \Psi$).

We take $\mathrm{P}(C)$ to be the number about which the relative frequency f/n of the event C stabilizes (i.e., the limit). This suggests some properties we want the function $\mathrm{P}(\cdot)$ to possess:

1. Relative frequency is always ≥ 0.
2. Relative frequency of S is 1.

3. If $C_1, C_2, C_3, \ldots$ are disjoint subsets of S and elements of Ψ, then the relative frequency of the union is the sum of the relative frequencies. This is called the *additive property.*

We can formalize these three properties mathematically:

Definition 2.3.1 (Probability Function). A *probability function* or *probability measure* is a set function $\mathrm{P}(\cdot)$ whose domain is Ψ. For every $a \in \Psi$, $\mathrm{P}(A)$ is called the *probability of the event A*. The following three axioms are assumed:

Axiom 1: $\mathrm{P}(A) \geq 0 \quad \forall A \in \Psi$.

Axiom 2: $\mathrm{P}(S) = 1$.

Axiom 3: If $A_1, A_2, A_3, \ldots \in \Psi$ are disjoint, then $\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathrm{P}(A_1 \cup A_2 \cup A_3 \cup \cdots) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$.

Note that the range of $\mathrm{P}(\cdot)$ is $[0, 1]$.

Example 2.3.2. Suppose we toss a fair die. Then $S = \{1, 2, 3, 4, 5, 6\}$, and for probability, we have

x	1	2	3	4	5	6
$\mathrm{P}(\{x\})$	1/6	1/6	1/6	1/6	1/6	1/6

Let A be the event that an even number is rolled:

$$A = \{2, 4, 6\} \qquad \Rightarrow \qquad \mathrm{P}(A) = \mathrm{P}(\{2\} \cup \{4\} \cup \{6\}) = \mathrm{P}(\{2\}) + \mathrm{P}(\{4\}) + \mathrm{P}(\{6\}) = \frac{3}{6} = \frac{1}{2}$$

□

We will henceforth assume that any event under discussion is an element of Ψ, the domain of $\mathrm{P}(\cdot)$.

It can be shown that it is possible to define the probability over the empty set $\varnothing$:

Theorem 2.3.1. $\mathrm{P}(\varnothing) = 0$.

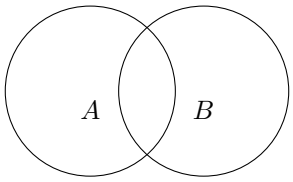
Proof:

- Since $S \cap \varnothing = \varnothing$, S and $\varnothing$ are disjoint. By axiom 3, $\mathrm{P}(S \cup \varnothing) = \mathrm{P}(S) + \mathrm{P}(\varnothing)$.
- Since $S \cup \varnothing = S$, $\mathrm{P}(S \cup \varnothing) = \mathrm{P}(S)$. By axiom 2, $\mathrm{P}(S) = 1$.
- By the above two steps, $1 = \mathrm{P}(S) = \mathrm{P}(S \cup \varnothing) = \mathrm{P}(S) + \mathrm{P}(\varnothing) = 1 + \mathrm{P}(\varnothing) \quad \Rightarrow \quad \mathrm{P}(\varnothing) = 1 - 1 = 0$.

■

Theorem 2.3.2. For any two events A and B, $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(AB)$.

Proof:



Note that

- $A \cup B = A \cup (B/AB)$, where $A \cap (B/AB) = \varnothing$.
- $B = (B/AB) \cup AB$, where $(B/AB) \cap AB = \varnothing$.

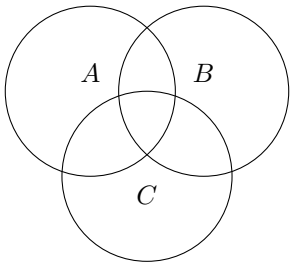
By axiom 3, $\mathrm{P}(A \cup B) = \mathrm{P}(A \cup (B/AB)) = \mathrm{P}(A) + \mathrm{P}(B/AB)$ and $\mathrm{P}(B) = \mathrm{P}((B/AB) \cup AB) = \mathrm{P}(B/AB) + \mathrm{P}(AB)$. Putting these two equations together proves the theorem.

■

Theorem 2.3.3. For any three sets A, B and C,

$$\mathrm{P}(A \cup B \cup C) = \mathrm{P}(A) + \mathrm{P}(B) + \mathrm{P}(C) - \mathrm{P}(AB) - \mathrm{P}(BC) - \mathrm{P}(AC) + \mathrm{P}(ABC).$$

Proof:



We see that $A \cup B \cup C = (A/AC) \cup (B/AB) \cup (C/BC) \cup ABC$, where all four sets are mutually disjoint. Also, $A = (A/AC) \cup AC$, etc. ■

Theorem 2.3.4. For any two events A and B, $\mathrm{P}(A \cup B) \leq \mathrm{P}(A) + \mathrm{P}(B)$.

Proof: By Theorem 2.3.2, $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(AB)$. By axiom 1, $\mathrm{P}(AB) \geq 0$. The conclusion directly follows. ■

Theorem 2.3.5. For any two events A and B, if $A \subset B$, then $\mathrm{P}(A) \leq \mathrm{P}(B)$.

Proof: We know that $B = (B/AB) \cup AB$, where $(B/AB) \cap AB = \varnothing$. Since $A \subset B$, then $AB = A$. By axiom 3, $\mathrm{P}(B) = \mathrm{P}(B/AB) + \mathrm{P}(AB) = \mathrm{P}(B/AB) + \mathrm{P}(A)$. Since $\mathrm{P}(B/AB) \geq 0$ by axiom 1, the result follows. ■

Theorem 2.3.6. For any event A, $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$.

Proof: $A \cup A^c = S$ and $A \cap A^c = \varnothing$. Thus, $1 = \mathrm{P}(S) = \mathrm{P}(A \cup A^c) = \mathrm{P}(A) + \mathrm{P}(A^c)$, so that $\mathrm{P}(A) = 1 - \mathrm{P}(A^c)$. ■

Example 2.3.3. As in Example 2.3.2, we again toss a fair die.

x	1	2	3	4	5	6
$\mathrm{P}(\{x\})$	1/6	1/6	1/6	1/6	1/6	1/6

Then

$$\begin{aligned} A = \{2,4,6\} &\Rightarrow \mathrm{P}(A) = \frac{3}{6} = \frac{1}{2}, \\ B = \{1,2,3\} &\Rightarrow \mathrm{P}(B) = \frac{3}{6} = \frac{1}{2}, \\ AB = \{2\} &\Rightarrow \mathrm{P}(AB) = \frac{1}{6}, \\ A \cup B = \{1,2,3,4,6\} &\Rightarrow \mathrm{P}(A \cup B) = \frac{5}{6}, \\ (AB)^c = \{1,3,4,5,6\} &\Rightarrow \mathrm{P}((AB)^c) = \frac{5}{6}. \end{aligned}$$

□

2.4 Combinatorics - Counting Techniques

Definition 2.4.1. An arrangement of n symbols in a definite order is called a *permutation* of those n symbols.

For example, $2,3,1,4$ is a permutation of the symbols $1,2,3,4$. The number of permutations of n things taken n at a time is

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1.$$

Note that we define $0! = 1$. The number of permutations of n items taken r at a time is

$${}_n\mathrm{P}_r = \frac{n!}{(n-r)!} = n(n-1)(n-2)\cdots(n-r+1).$$

Example 2.4.1. The number of permutations of a, b, c taken three at a time is $3! = 6$:

$$abc \quad bac \quad cab \quad acb \quad bca \quad cba.$$

The number of permutations of a, b, c taken two at a time is ${}_3\mathrm{P}_2 = 6$:

$$ab \quad ba \quad ca \quad ac \quad bc \quad cb.$$

The number of permutations of a, b, c and d taken two at a time is ${}_4\mathrm{P}_2 = 12$:

$$ab \quad ba \quad ca \quad da \quad ac \quad bc \quad cb \quad db \quad ad \quad bd \quad cd \quad dc.$$

□

Example 2.4.2. If *repeats* are possible, the number of ways that the symbols a and b can be arranged in a definite order of 2 symbols is $2 \cdot 2 = 4$:

$$aa \quad ba \quad ab \quad bb.$$

The three symbols a, b and c can be arranged in $3 \cdot 3 \cdot 3 = 27$ ways if repeats are allowed. □

Example 2.4.3. Consider an urn with 1 white and 2 black balls. Number the white ball #1 and the black balls #2 and #3. Draw two balls from an urn *without replacement*. Record the outcome by (x_1, x_2), where $x_1, x_2 \in \{1, 2, 3\}$ but $x_1 \neq x_2$. (x_i denotes the result of the i^{th} draw):

$$S = \{(x_1, x_2) : \ x_1, x_2 \in \{1, 2, 3\}; \ x_1 \neq x_2\}.$$

How many possible outcomes are there? That is, what is the size of the sample space?

$${}_3\mathrm{P}_2 = 3 \cdot 2 = \frac{3!}{(3-2)!} = 6, \qquad \Rightarrow \qquad \begin{matrix} (1,2), & (2,1), & (3,1), \\ (1,3), & (2,3) & (3,2). \end{matrix}$$ □

In general, the number of r-tuples that can be formed from n symbols when we don't repeat is ${}_n\mathrm{P}_r$. If repetition is allowed, then the number is n^r. In Example 2.4.3, if the two balls are drawn from the urn *with replacement*, then the sample space becomes

$$S = \{(x_1, x_2) : \ x_1, x_2 \in \{1, 2, 3\}\}.$$

and the number of possible outcomes is

$$3 \cdot 3 = 3^2 = n^r = 9.$$

Definition 2.4.2. The number of subsets of size r that can be formed from a set of n elements is called a *combination* of n elements taken r at a time, denoted by

$$\binom{n}{r} = \frac{n!}{(n-r)!\, r!} = \frac{{}_n\mathrm{P}_r}{r!}.$$

Example 2.4.4. Consider the set $\{1, 2, 3\}$. The number of combinations on this set taken two at a time is

$$\binom{3}{2} = \frac{3!}{2!1!} = \frac{{}_3\mathrm{P}_2}{2!} = 3, \qquad \Rightarrow \qquad \{1,2\}, \quad \{1,3\}, \quad \{2,3\}.$$

As illustrated in Example 2.4.3, the number of permutations on this set taken two at a time is

$${}_3\mathrm{P}_2 = 3 \cdot 2 = \frac{3!}{(3-2)!} = 6, \qquad \Rightarrow \qquad \begin{matrix} (1,2), & (2,1), & (3,1), \\ (1,3), & (2,3) & (3,2). \end{matrix}$$ □

Example 2.4.5. The number of different bridge hands (13 cards) that a player in a bridge game can attain is

$$\binom{52}{13} = \frac{52!}{39!13!} = 635,013,559,600 \approx 6.35 \times 10^{11}.$$

The number of ways in which a bridge deck may be dealt into four hands for four distinct players is

$$\binom{52}{13}\binom{39}{13}\binom{26}{13}\binom{13}{13} = \frac{52!}{\cancel{39!}13!} \cdot \frac{\cancel{39!}}{\cancel{26!}13!} \cdot \frac{\cancel{26!}}{\cancel{13!}13!} \cdot \frac{\cancel{13!}}{0!13!} = \frac{52!}{13!13!13!13!} \approx 5.36 \times 10^{28}.$$ □

The answer to the last part of Example 2.4.5 has a special notation:

$$\binom{n}{r_1, r_2, \ldots, r_k} = \frac{n!}{r_1!\ r_2!\ \cdots\ r_k!}, \qquad \text{where } r_1 + \cdots + r_k = n.$$

Therefore, we could have denoted the answer to the last part of that example as

$$\binom{52}{13, 13, 13, 13} = \frac{52!}{13!13!13!13!} \approx 5.36 \times 10^{28}.$$

Example 2.4.6. Five probability students meet at a party. How many handshakes are exchanged if each student shakes hands with every other student once and only once? To determine this, we simply calculate the total combinations of two people (the two people shaking hands) out of a total of five:

$$\binom{5}{2} = \frac{5!}{2!3!} = \frac{5 \cdot 4}{2 \cdot 1} = 10. \qquad \square$$

Example 2.4.7. In how many ways can a probability student answer 8 questions on a true-false exam if she makes half the questions true and half the questions false? Here we take 8 questions and choose 4 of them to be false:

$$\binom{8}{4} = \frac{8!}{4!4!} = \frac{8 \cdot 7 \cdot 6 \cdot 5}{4 \cdot 3 \cdot 2 \cdot 1} = 70. \qquad \square$$

2.5 Defining P when S is Finite

To define P over a finite sample space S, it is necessary to know P for any subset of S. It is sufficient to define P over the single element events. That is, if $S = \{x_1, \ldots, x_n\}$ and if we know $\mathrm{P}(x_1)$, $\ldots$, $\mathrm{P}(x_n)$, then we can compute $\mathrm{P}(\cdot)$ for any subset of S.

Example 2.5.1. Suppose one is drawing a sample of size 2 from an urn containing red and white balls:

$$S = \{(W,W), (W,R), (R,W), (R,R)\}.$$

Suppose we define the following.

x	(W,W)	(W,R)	(R,W)	(R,R)
$\mathrm{P}(\{x\})$	6/15	4/15	4/15	1/15

Let E be the event that the first ball is white:

$$\begin{aligned} E = \{(W,W), (W,R)\} \quad \Rightarrow \quad \mathrm{P}(E) &= \mathrm{P}\left(\{(W,W), (W,R)\}\right) \\ &= \mathrm{P}\left(\{(W,W)\} \cup \{(W,R)\}\right) \\ &= \mathrm{P}\left(\{(W,W)\}\right) + \mathrm{P}\left(\{(W,R)\}\right) \\ &= \frac{6}{15} + \frac{4}{15} \\ &= \frac{2}{3}. \end{aligned}$$

Note that the points (single-element events) in S are not equally likely. $\square$

Now consider a *finite sample space*, where each point in the sample space is assumed to be equally likely:

$$S = \{x_1, x_2, \ldots, x_n\} \qquad \Rightarrow \qquad \mathrm{P}(\{x_i\}) = \frac{1}{n} \quad \text{for } i \in \{1, 2, \ldots, n\}.$$

In this special case, $\mathrm{P}(A) = \frac{\text{Size of } A}{\text{Size of } S}$. Therefore, if there are m points in A, then

$$\mathrm{P}(A) = \sum_{i=1}^{m} \frac{1}{n} = \frac{m}{n}.$$

Example 2.5.2. Consider the game of tossing 2 dice, where

$$S = \{(x, y) : \ x, y \in \{1, 2, 3, 4, 5, 6\}\}.$$

If it is assumed that each point in S is equally likely, then

$$\mathrm{P}(\{(x, y)\}) = \frac{1}{\text{Size of } S} \quad \forall (x, y) \in S.$$

Let A be the set of points such that $x + y = 5$:

$$A = \{(1, 4), (4, 1), (2, 3), (3, 2)\} \qquad \Rightarrow \qquad \mathrm{P}(A) = \frac{\text{Size of } A}{\text{Size of } S} = \frac{4}{6^2} = \frac{4}{36} = \frac{1}{9}.$$

Let $S = \{2, 3, 4, 5, \ldots, 11, 12\}$, where we are now *only* interested in the events concerning the sum $x + y$ appearing on the dice. *The points in S are not equally likely.*

x	2	3	4	5	6	7	8	9	10	11	12
$\mathrm{P}(\{x\})$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

If the event $A = \{5\}$, then $\mathrm{P}(A) = \frac{4}{36} = \frac{1}{9}$. □

Example 2.5.3. Find the probability that the thirteenth day of a randomly chosen month (each month equally likely to be chosen) is Friday:

$$S = \{\text{Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday}\}.$$

If we assume the points in S are equally likely, then $\mathrm{P}(\text{Friday}) = \frac{1}{7}$. Checking the assumption of equally likely points: A calendar has a period of 400 years. Every fourth year is a leap year except for years such as 1700, 1800, and 1900, which are not multiples of 400. In 400 years, there are 4800 dates (between the years 1600 and 2000) which are the thirteenth of the month:

x	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
$\mathrm{P}(\{x\})$	$\frac{687}{4800}$	$\frac{685}{4800}$	$\frac{685}{4800}$	$\frac{687}{4800}$	$\frac{684}{4800}$	$\frac{688}{4800}$	$\frac{684}{4800}$

□

Example 2.5.4 (Urn Problems). An urn contains 20 balls:

10 red, numbered 1-10,
4 blue, numbered 11-14,
6 green, numbered 15-20.

(a) A sample of size 1 is taken from the urn.

$$S = \{1, 2, \ldots, 20\} \quad \Rightarrow \quad \text{Size of } S \text{ is } 20.$$

Assuming equally likely points in S, let A be the event that the ball drawn is red.

$$\mathrm{P}(A) = \mathrm{P}(\{1, 2, \ldots, 10\}) = \frac{\text{Size of } A}{\text{Size of } S} = \frac{10}{20} = \frac{1}{2}.$$

(b) A sample of size 2 is drawn *with replacement* from the urn.

$$S = \{(x, y) : \ x, y \in \{1, 2, \ldots, 20\}\} \quad \Rightarrow \quad \text{Size of } S \text{ is } 20 \times 20 = 400.$$

Let B be the event that the first ball is red and the second is green.

$$\begin{aligned} B = \{(x, y) : \ x \in \{1, 2, \ldots, 10\}, \ y \in \{15, 16, \ldots, 20\}\} \quad &\Rightarrow \quad \text{Size of } B \text{ is } 10 \times 6 = 60. \\ &\Rightarrow \quad \mathrm{P}(B) = \frac{\text{Size of } B}{\text{Size of } S} = \frac{60}{400} = \frac{3}{20}. \end{aligned}$$

Let C be the event that the one ball is red and one ball is green.

$$\begin{aligned} C = \{(x, y) : \ x \in \{1, \ldots, 10\}, \, y \in \{15, \ldots, 20\} \text{ or } x \in \{15, \ldots, 20\}, \ y \in \{1, \ldots, 10\}\} \\ \Rightarrow \quad \text{Size of } C \text{ is } 10 \times 6 + 6 \times 10 = 120. \\ \Rightarrow \quad \mathrm{P}(C) = \frac{\text{Size of } C}{\text{Size of } S} = \frac{120}{400} = \frac{3}{10}. \end{aligned}$$

Let D be the event that both balls are red.

$$\begin{aligned} D = \{(x, y) : \ x, y \in \{1, 2, \ldots, 10\}\} \quad &\Rightarrow \quad \text{Size of } D \text{ is } 10 \times 10 = 100. \\ &\Rightarrow \quad \mathrm{P}(D) = \frac{\text{Size of } D}{\text{Size of } S} = \frac{100}{400} = \frac{1}{4}. \end{aligned}$$

(c) A sample of size 2 is drawn *without replacement* from the urn.

$$S = \{(x, y) : \ x, y \in \{1, 2, \ldots, 20\}, \ x \neq y\} \quad \Rightarrow \quad \text{Size of } S \text{ is } 20 \times 19 = 380.$$

Assuming equally likely points in S, let E be the event that the first ball drawn is red and the second one is green.

$$\begin{aligned} E = \{(x, y) : \ x \in \{1, 2, \ldots, 10\}, \ y \in \{15, 16, \ldots, 20\}\} \quad &\Rightarrow \quad \text{Size of } E \text{ is } 10 \times 6 = 60. \\ &\Rightarrow \quad \mathrm{P}(E) = \frac{\text{Size of } E}{\text{Size of } S} = \frac{60}{380} > \mathrm{P}(B) = \frac{60}{400}. \end{aligned}$$

Note that if there are 19 balls in the urn and 6 of them are green, then the probability of a green ball is $\frac{6}{19} > \frac{6}{20}$. Now let F be the event that both balls drawn are red.

$$\begin{aligned} F = \{(x, y) : \ x, y \in \{1, 2, \ldots, 10\}, \ x \neq y\} \quad &\Rightarrow \quad \text{Size of } F \text{ is } 10 \times 9 = 90. \\ &\Rightarrow \quad \mathrm{P}(F) = \frac{\text{Size of } F}{\text{Size of } S} = \frac{90}{380} < \mathrm{P}(D) = \frac{100}{400}. \end{aligned}$$

□

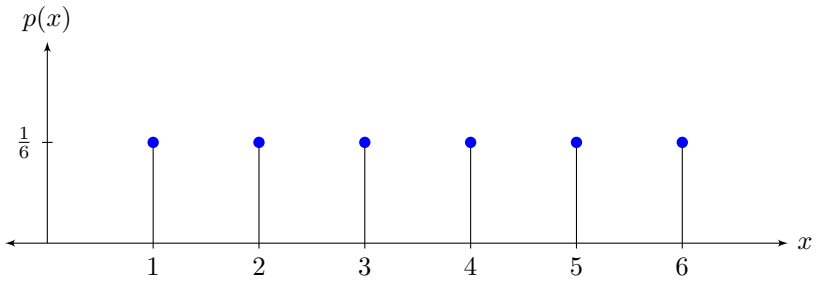
When S is countably infinite, it is obvious that the points in S cannot be equally likely. To define $\mathrm{P}(\cdot)$ in this case, it is sufficient to define $\mathrm{P}(\cdot)$ over the single member equivalents of S.

Definition 2.5.1 (Probability Mass/Density Function). For finite or countably infinite sample spaces, the probability function defined over single member events of S is written as a point function and is called the *probability mass function* (pmf) or *probability density function* (pdf):

$$\begin{aligned} \text{pmf: } & p(x) = \mathrm{P}(\{x\}), \\ \text{pdf: } & f(x) = \mathrm{P}(\{x\}). \end{aligned}$$

Example 2.5.5. Toss a fair 6-sided die:

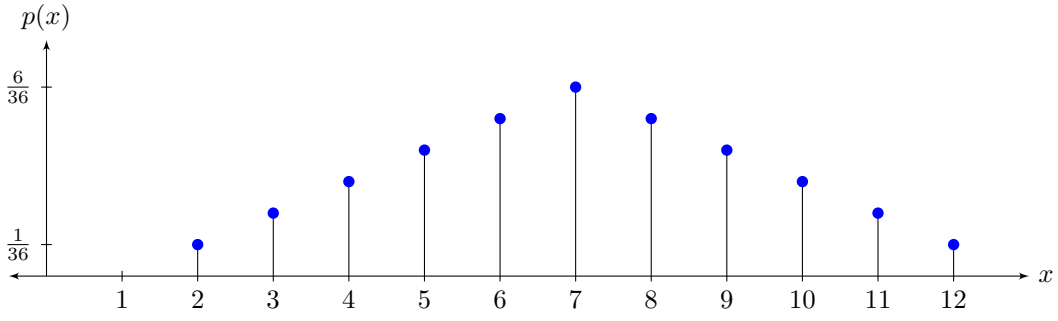
$$p(x) = \mathrm{P}(\{x\}) = \begin{cases} \frac{1}{6} & x \in \{1,2,3,4,5,6\} \\ 0 & \text{otherwise} \end{cases}.$$



□

Example 2.5.6. Toss 2 fair dice as in Example 2.5.2. Let the outcome by the sum $x+y$ of the values of their faces:

$$p(x) = \begin{cases} \frac{x-1}{36} & x \in \{2,3,4,5,6,7\} \\ \frac{13-x}{36} & x \in \{8,9,10,11,12\} \\ 0 & \text{otherwise} \end{cases}.$$



□

Example 2.5.7. Toss a coin, and let p be the probability of getting a head. Toss a coin over and over again until a head appears and stop. We are interested in the number of tosses, x, necessary to obtain a head for the first time.

$$S = \{1, 2, 3, \ldots\} = \mathbb{Z}^+.$$

It can be shown that

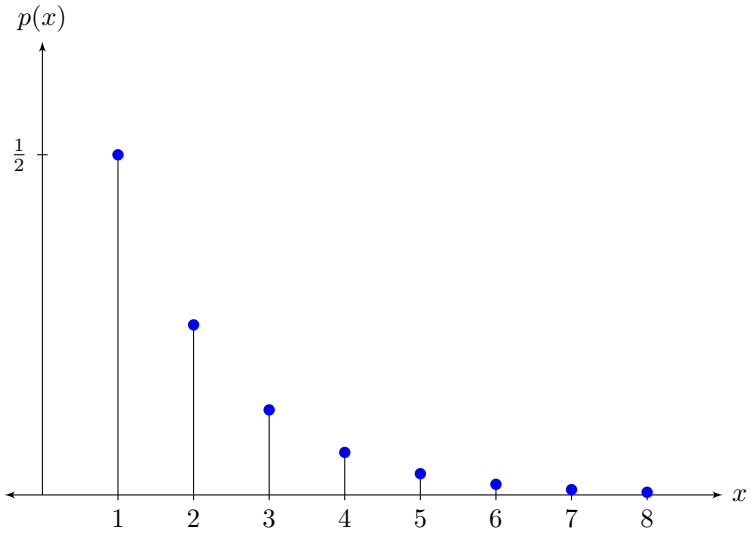
$$\begin{aligned} p(1) &= p \\ p(2) &= (1-p)p \\ p(3) &= (1-p)^2 p \\ &\vdots \\ p(x) &= (1-p)^{x-1} p. \end{aligned}$$

Therefore, we write

$$p(x) = \begin{cases} (1-p)^{x-1} p & x \in \mathbb{Z}^+ \\ 0 & \text{otherwise} \end{cases}.$$

If $p = \frac{1}{2}$, then

$$p(x) = \begin{cases} \left(\frac{1}{2}\right)^x & x \in \mathbb{Z}^+ \\ 0 & \text{otherwise} \end{cases}.$$



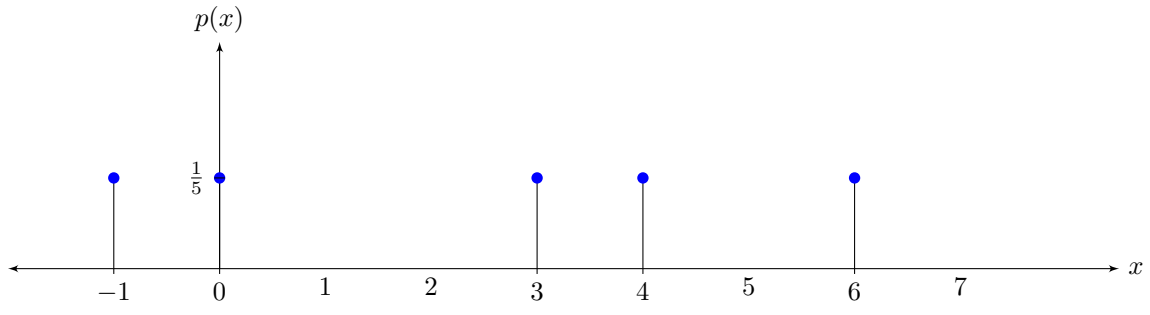
□

If S has n members and each is equally likely to occur, then we say that $\mathrm{P}(\cdot)$ is *distributed uniformly over* S, and the pmf is

$$p(x) = \begin{cases} \frac{1}{n} & x \in S \\ 0 & \text{otherwise} \end{cases}.$$

Example 2.5.8. If $S = \{-1, 0, 3, 4, 6\}$ and $\mathrm{P}(\cdot)$ is distributed uniformly over S, then

$$p(x) = \begin{cases} \frac{1}{5} & x \in \{-1, 0, 3, 4, 6\} \\ 0 & \text{otherwise} \end{cases}.$$



□

2.6 Defining P when S is (Uncountably) Infinite

We will consider cases where $S = \mathbb{R}$. We will restrict our attention to those sample spaces S and set of events Ψ which we can define probability P for any event $A \in \Psi$ in the following way:

If there exists a real-valued function $f(\cdot)$ such that

$$\mathrm{P}(A) = \int_A f(x)\, dx \quad \forall A \in \Psi, \tag{2.6.1}$$

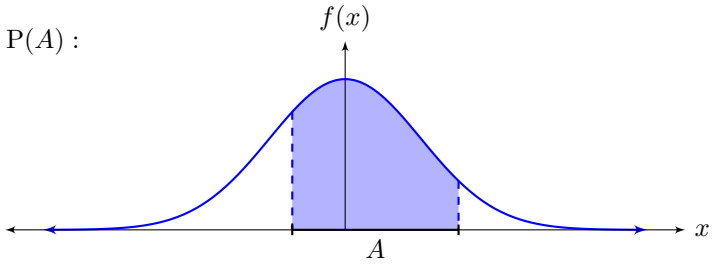
$$1 = \mathrm{P}(S) = \int_S f(x)\, dx, \text{and} \tag{2.6.2}$$

$$f(x) \geq 0 \ \ \forall x \in S \subset \mathbb{R}, \tag{2.6.3}$$

then $f(\cdot)$ is called a *probability density function*. Note that

$$\mathrm{P}(\{x_i\}) = \int_{x_i}^{x_i} f(x)\, dx = 0.$$

However, $\mathrm{P}(\cdot)$ is still a set function.



Example 2.6.1. Assume that the subway station at T-Centralen has trains which depart every ten minutes for Gamla Stan station. Passengers arriving at T-Centralen don't know the exact schedule – only that there is a ten-minute interval between trains. The waiting time of a passenger arriving at T-Centralen is t, where $t \in [0, 10]$:

$$S = \{t : 0 \leq t \leq 10\}.$$

Experience has shown that for $a, b \in [0, 10]$,

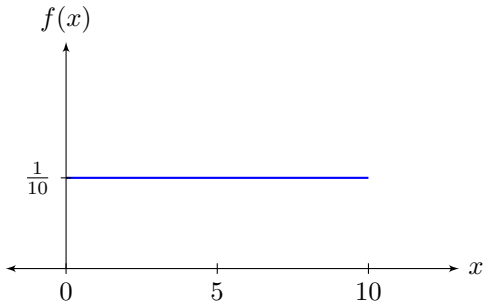
$$\mathrm{P}(\{t : a \leq t \leq b\}) = k(b - a) = \int_a^b k \, dx.$$

By (2.6.2),

$$\mathrm{P}(\{t : 0 \leq t \leq 10\}) = \mathrm{P}(S) = 1 = k(10 - 0) = 10k \quad \Rightarrow \quad k = \frac{1}{10}.$$

Therefore,

$$f(x) = \begin{cases} \frac{1}{10} & x \in [0, 10] \\ 0 & \text{otherwise} \end{cases}.$$

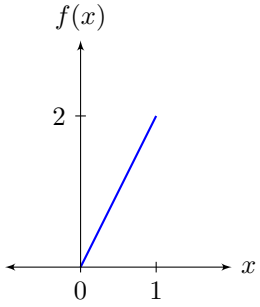


so that

$$\mathrm{P}(\{t : a \leq t \leq b\}) = \int_a^b \frac{1}{10} \, dx = \frac{b - a}{10} \quad \text{for } a, b \in [0, 10]. \qquad \square$$

Example 2.6.2. Let $S = \mathbb{R}$, and

$$f(x) = \begin{cases} 2x & x \in [0, 1] \\ 0 & \text{otherwise} \end{cases}.$$



Note that

$$\int_{-\infty}^{\infty} f(x) \, dx = \int_0^1 2x \, dx = \left[x^2\right]_0^1 = 1.$$

Let A be the event that the observed value is less than $\frac{1}{2}$. Thus,

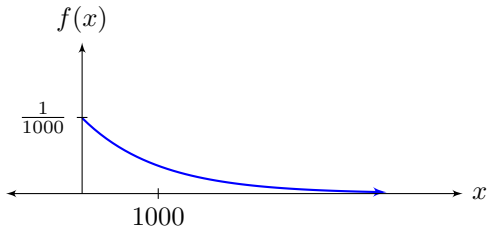
$$\mathrm{P}(A) = \int_{-\infty}^{\frac{1}{2}} f(x) \, dx = \int_{-\infty}^{\frac{1}{2}} 2x \, dx = \left[x^2\right]_0^{\frac{1}{2}} = \frac{1}{4}.$$

and

$$\mathrm{P}(\text{observed value is} \geq \tfrac{1}{2}) = 1 - \frac{1}{4} = \frac{3}{4}, \qquad \mathrm{P}(\{\tfrac{1}{2}\}) = \int_{\frac{1}{2}}^{\frac{1}{2}} 2x \, dx = 0. \qquad \square$$

Example 2.6.3. Consider an experiment which consists of observing the total time a light bulb will burn from the moment it is first put into service. Suppose $\mathrm{P}(\cdot)$ is determined by the pdf

$$f(x) = \begin{cases} \frac{1}{1000} e^{-x/1000} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$



Let E be the event that the bulb burns between 100 and 1000 hours. Let F be the event that the bulb burns more than 1000 hours. That is,

$$E = \{x : \ 100 \leq x \leq 1000\}, \quad F = \{x : \ x > 1000\}.$$

Thus,

$$\mathrm{P}(E) = \int_{100}^{1000} f(x)\ dx = \int_{100}^{1000} \frac{1}{1000} e^{-x/1000}\ dx = \left[-e^{-x/1000}\right]_{100}^{1000} = -e^{-1} + e^{-\frac{1}{10}} \approx 0.537,$$
$$\mathrm{P}(F) = \int_{1000}^{\infty} f(x)\ dx = \int_{1000}^{\infty} \frac{1}{1000} e^{-x/1000}\ dx = \left[-e^{-x/1000}\right]_{1000}^{\infty} = 0 + e^{-1} \approx 0.368.$$

Furthermore, note that $\mathrm{P}(\{x : \ x < 100\}) = 1 - \mathrm{P}(E) - \mathrm{P}(F) \approx 1 - 0.537 - 0.368 = 0.095$ and that

$$\int_{-\infty}^{\infty} f(x)\ dx = \int_{0}^{\infty} \frac{1}{1000} e^{-x/1000}\ dx = \left[-e^{-x/1000}\right]_{0}^{\infty} = 0 + 1 = 1,$$

thus proving that $f(x)$ is a valid pmf (i.e., it integrates to one). □

2.7 Random Variables

The set of all outcomes of a random[1] experiment is the sample space for the random experiment. Examples of random experiments include the following:

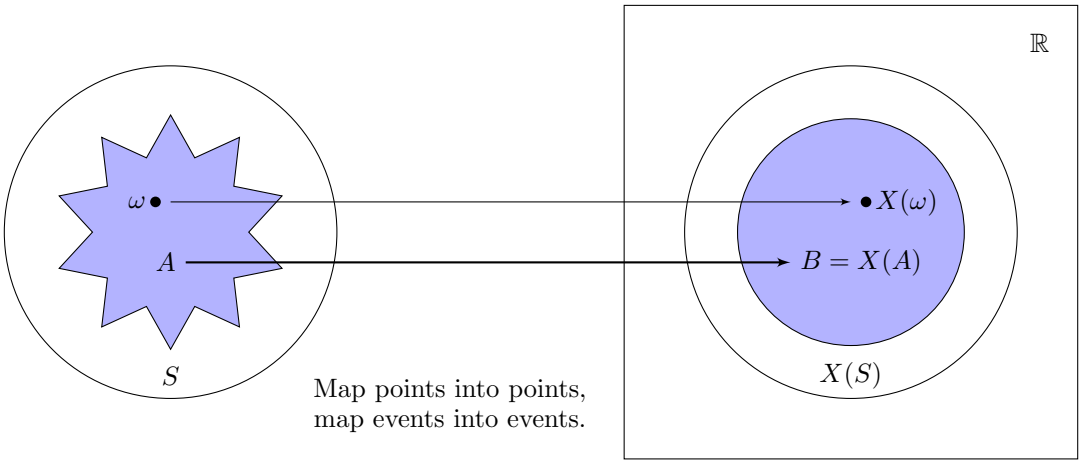
- Measuring the time a passenger waits until he catches a train.
- Measuring the time to failure of a piece of equipment.
- Observing the sum when two dice are tossed.

For these examples, we could let

- T denote the time a passenger waits,
- X denote the time to failure, and
- Y denote the sum of the two dice.

[1] The adjective (or adverb) *random* is used in probability in many different applications, such as random variable, random phenomenon, random experiment, and picking randomly. In all cases, this just means that the outcome is not pre-determined, and relies on some law of chance.

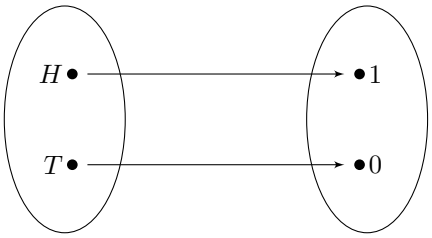
Definition 2.7.1 (Random Variable). A *random variable* (r.v.) X is a point function from the sample space S into the set of real numbers $\mathbb{R}$. If B is a probabilizable subset of $\mathbb{R}$, then $\mathrm{P}(X \in B) = \mathrm{P}_X(B) = \mathrm{P}(\{\omega : \ X(\omega) \in B\})$. We let $A = X^{-1}(B)$ be the *inverse image of B in S.*



We denote random variables by upper case letters such as X, Y, Z or T. We denote specific values which the random variables may take by lower case letters, such as x, y, z or t.

Example 2.7.1. Suppose $S = \{1, 2, 3, 4, 5, 6\}$ and X is the *identity function* $X(\omega) = \omega \ \forall \omega \in S$. Thus, for example, $X(1) = 1$, $X(2) = 2$, etc. If $\mathrm{P}(x) = \frac{1}{6}$ for each value x of S, then $\mathrm{P}(X = x) = \frac{1}{6} \ \forall x$. □

Example 2.7.2. If $S = \{H, T\}$, then we could let $X(H) = 1$, $X(T) = 0$. By doing this, we are mapping the sample space $\{H, T\}$ to the subset of the real numbers $\{0, 1\}$.



$$\mathrm{P}(\{H\}) = \mathrm{P}(X = 1),$$
$$\mathrm{P}(\{T\}) = \mathrm{P}(X = 0).$$

□

Example 2.7.3. Let $S = \mathbb{R}$, and let E be the event that a light bulb burns between 100 and 1000 hours. Letting $X(\omega) = \omega \ \forall \omega \in S$, we have $\mathrm{P}(E) = \mathrm{P}(\{x : \ 100 < x < 1000\}] = \mathrm{P}(100 < X < 1000)$. □

We will learn two types of random variables: *discrete* and *continuous.*

Definition 2.7.2. A random variable X is *discrete* iff its range of values (values of x where $\mathrm{P}(X = x) > 0$) is a non-empty finite or countably infinite set of real numbers.

Example 2.7.4. The random variables in Examples 2.7.1 (die) and 2.7.2 (coin) are discrete. The random variable in Example 2.7.3 (light bulb) is not discrete, since the set of x where $\mathrm{P}(X = x) > 0$ is the empty set $\varnothing$. □

If S is finite or countably infinite and $X(x) = x \ \forall x \in S$, then X is discrete, such as in Example 2.7.1. However, discrete random variables exist in other cases, such as in Examples 2.7.2 and 2.7.5 below.

Example 2.7.5.

a. Let $S = \mathbb{Z}^+$, the set of natural numbers. Define $X(x) = \begin{cases} -x & \text{if } x = 2n+1 \text{ for some } n \in \mathbb{Z}^+, \\ x/2 & \text{if } x = 2n \text{ for some } n \in \mathbb{Z}^+ \end{cases}$. Then X is a discrete random variable.

b. If we define $X(x) = \begin{cases} 0 & \text{if } x = 2n+1 \text{ for some } n \in \mathbb{Z}^+, \\ 1 & \text{if } x = 2n \text{ for some } n \in \mathbb{Z}^+ \end{cases}$, then X is a discrete random variable.

c. Let $S = \mathbb{R}^+$, and define $X(x) = \begin{cases} 0 & x \in (0, 100) \\ 1 & x \geq 100 \end{cases}$. Then X is a discrete random variable.

□

Definition 2.7.3. A random variable X is *continuous* iff its range is an uncountably infinite set of real numbers. For example, $\forall x \in \mathbb{R},\ \mathrm{P}(X = x) = 0$.

We will restrict our attention to those continuous r.v.'s where $\mathrm{P}(\cdot)$ is defined by a pdf $f_X(x)$. These random variables are correctly called *absolutely continuous* random variables.

If X is a continuous r.v., then $\mathrm{P}(X = x) = 0\ \forall x$ and $\forall A \in \Psi,\ \mathrm{P}(A) = \mathrm{P}(X \in A)$.

Example 2.7.6. Let $S = \{x :\ 0 \leq x \leq 10\}$ and $X(x) = x\ \forall x \in S$. Let $f_X(x) = \frac{1}{10}$ for $0 \leq x \leq 10$. then

$$\mathrm{P}(a < x < b) = \int_a^b \frac{1}{10}\, dx \quad \text{for } 0 < a < b < 10.$$

□

Example 2.7.7. Let $S = \mathbb{R}^+$, where S is a sample space for a random experiment of life length of light bulbs. Let

$$\mathrm{P}(A) = \int_A \frac{1}{100} e^{-x/100}\, dx \quad \forall A \in \Psi.$$

□

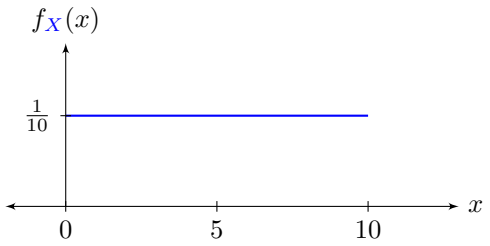
Example 2.7.8. Let $S = \{H, T\}$ and $\mathrm{P}(\{H\}) = \mathrm{P}(\{T\}) = \frac{1}{2}$. If $X(H) = 1$ and $X(T) = 0$, then for this discrete random variable.

$$p(1) = \mathrm{P}(\{X(H) = 1\}) = \mathrm{P}(\{H\}) = \mathrm{P}(X = 1),$$
$$p(0) = \mathrm{P}(\{X(T) = 0\}) = \mathrm{P}(\{T\}) = \mathrm{P}(X = 0).$$

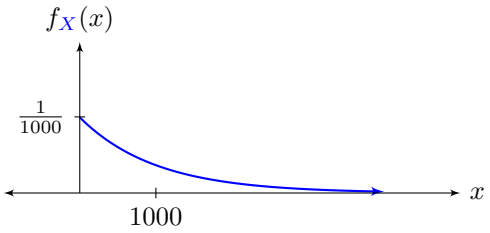
□

Example 2.7.9. Further examples of pmf's and pdf's include the following:

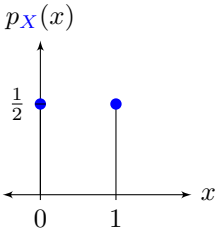
a) $f_X(x) = \begin{cases} \frac{1}{10} & x \in [0, 10] \\ 0 & \text{otherwise} \end{cases}$.



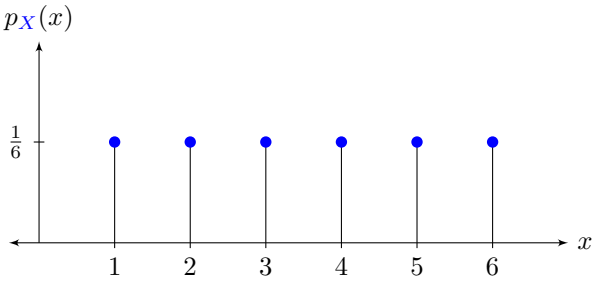
b) $f_X(x) = \begin{cases} \frac{1}{1000} e^{-x/1000} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$.



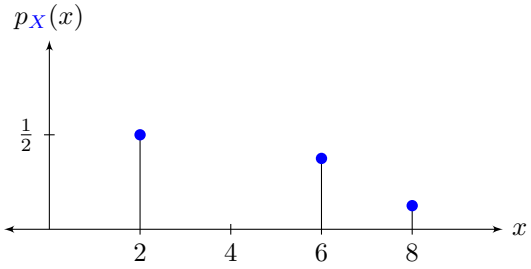
c) $p_X(x) = \begin{cases} \frac{1}{2} & x \in \{0, 1\} \\ 0 & \text{otherwise} \end{cases}$.



d) $p_X(x) = \begin{cases} \frac{1}{6} & x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise} \end{cases}$.



e) $p_X(x) = \begin{cases} \frac{1}{2} & x = 2 \\ \frac{3}{8} & x = 6 \\ \frac{1}{8} & x = 8 \\ 0 & \text{otherwise} \end{cases}$.



□

2.8 Distribution Function

It can be shown that for any random variable, there exists a function called a *distribution function* which suffices to determine the probability measure $\mathrm{P}(\cdot)$, in that $\mathrm{P}(\cdot)$ can be reconstructed from the distribution function.

Definition 2.8.1 (Distribution Function). For a random variable X, the *distribution function* $F_X(\cdot)$ is a point function from $\mathbb{R}$ to the interval $[0,1]$ defined by

$$F_X(x) = \mathrm{P}(\{y:\ y \le x\}) = \mathrm{P}(X \le x).$$

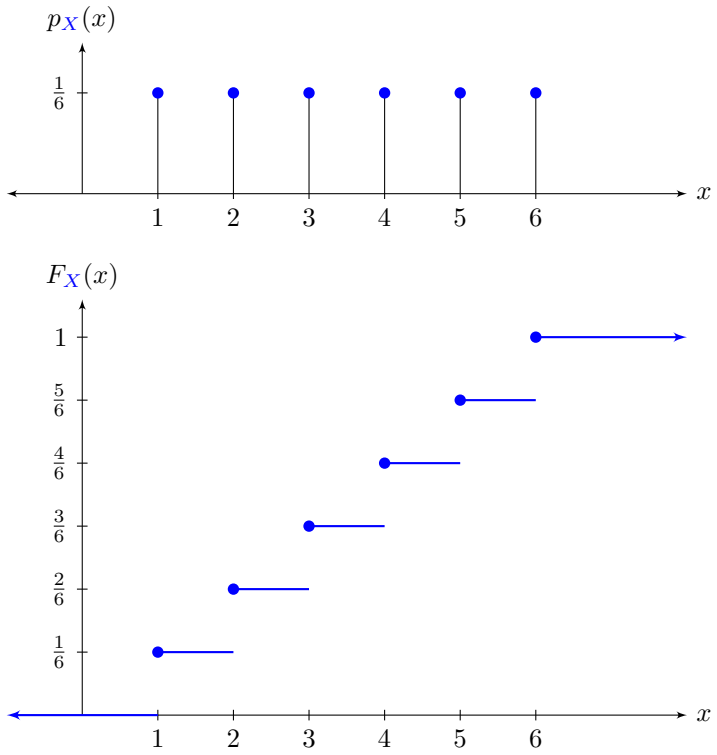
If the probability is specified by a pmf $p_X(\cdot)$ (for discrete r.v.'s), then $F_X(x) = \sum_{y \le x} p_X(y)$.

If the probability is specified by a pdf $f_X(\cdot)$ (for continuous r.v.'s), then $F_X(x) = \int_{-\infty}^{x} f_X(y)\ dy$.

Example 2.8.1. If X is a discrete r.v. with $S = \{1,2,3,4,5,6\}$ and

$$p_X(x) = \begin{cases} \frac{1}{6} & x \in S \\ 0 & \text{otherwise} \end{cases}, \quad \text{then} \quad F_X(x) = \begin{cases} 0 & x < 1 \\ \frac{1}{6} & 1 \le x < 2 \\ \frac{2}{6} & 2 \le x < 3 \\ \frac{3}{6} & 3 \le x < 4 \\ \frac{4}{6} & 4 \le x < 5 \\ \frac{5}{6} & 5 \le x < 6 \\ 1 & 6 \le x \end{cases}.$$

Graphically, we have

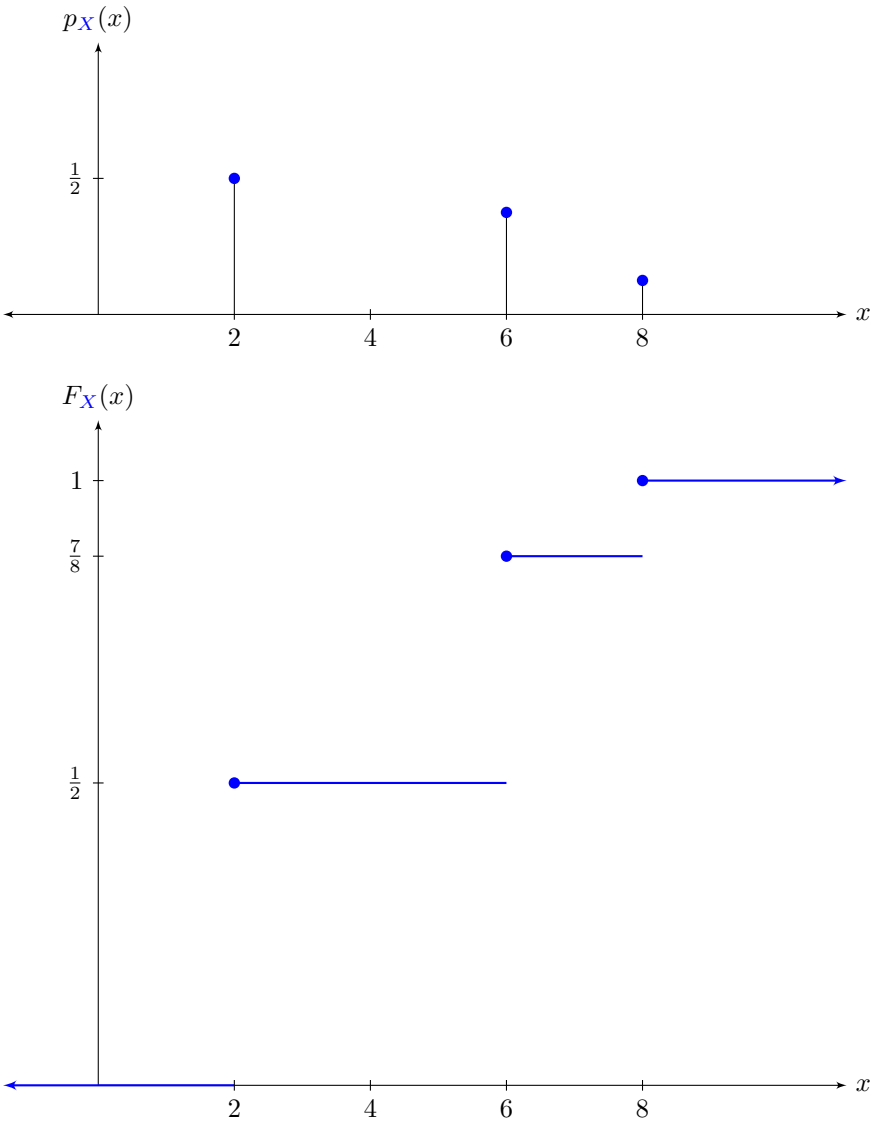


Note that the graph of $F_X(x)$ contains a line from $(1,0)$ to $(-\infty,0)$ (difficult to see on a black and white copy). □

Example 2.8.2. If X is a discrete r.v. with $S = \{2, 6, 8\}$ and

$$p_X(x) = \begin{cases} \frac{1}{2} & x = 2 \\ \frac{3}{8} & x = 6 \\ \frac{1}{8} & x = 8 \\ 0 & \text{otherwise} \end{cases}, \quad \text{then} \quad F_X(x) = \begin{cases} 0 & x < 2 \\ \frac{1}{2} & 2 \le x < 6 \\ \frac{7}{8} & 6 \le x < 8 \\ 1 & 8 \le x \end{cases}.$$

Graphically, we have



Here the graph of $F_X(x)$ contains a line from $(2, 0)$ to $(-\infty, 0)$ (again, difficult to see on a black and white copy).□

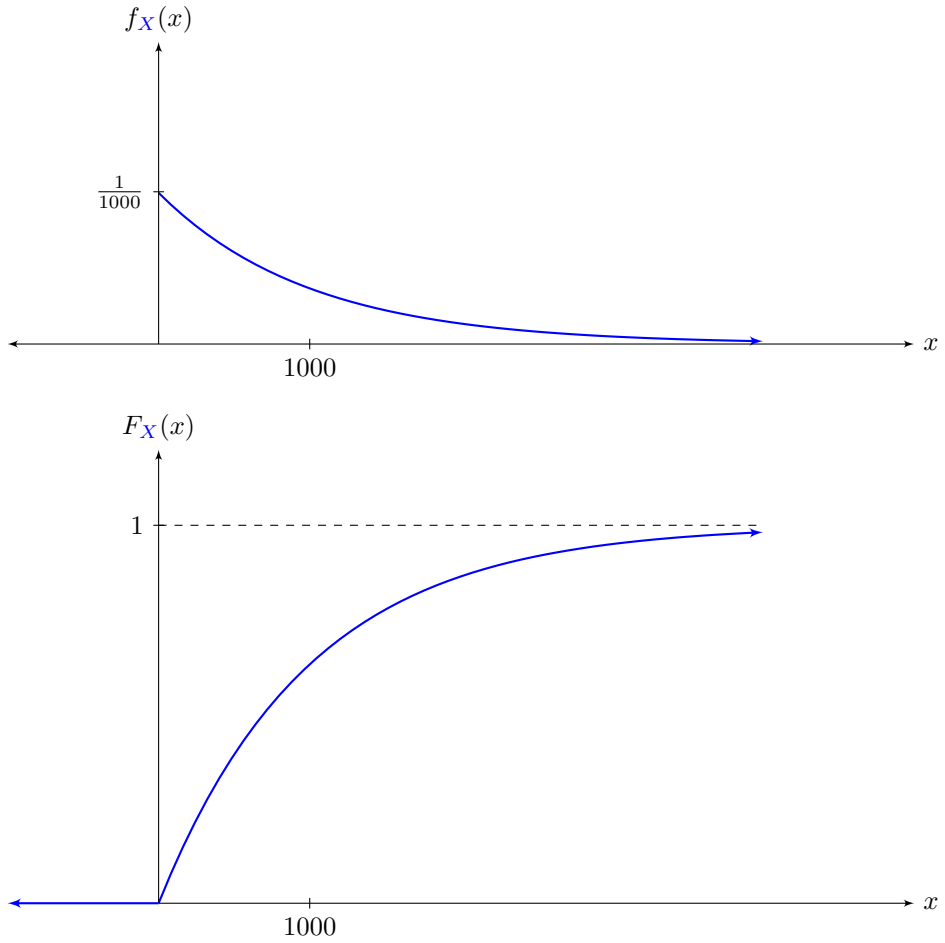
Example 2.8.3. Consider again the example of the life length of light bulbs:

$$f_X(x) = \begin{cases} \frac{1}{1000}e^{-x/1000} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

in which case, for $x \geq 0$,

$$F_X(x) = \mathrm{P}(X \leq x) = \int_0^x \frac{1}{1000}e^{-t/1000}\, dt = \left[-e^{-t/1000}\right]_0^x = 1 - e^{-x/1000}.$$

Graphically, we have



The graph of $F_X(x)$ contains a line from $(0, 0)$ to $(-\infty, 0)$ (again, difficult to see on a black and white copy). □

Generally, remember that $F_X(\cdot)$ is defined on the entire real number line $\mathbb{R}$. Furthermore, there is a 1-1 correspondence among

$$\mathrm{P}_X(\cdot) \quad \Leftrightarrow \quad F_X(\cdot) \quad \Leftrightarrow \quad p_X(\cdot)$$

for discrete r.v.'s and among

$$\mathrm{P}_X(\cdot) \quad \Leftrightarrow \quad F_X(\cdot) \quad \Leftrightarrow \quad f_X(\cdot)$$

for continuous r.v.'s.

2.9 n-Dimensional Discrete Random Variables

In this case the sample space is a discrete set of points in n-dimensional space. Let $S_1, S_2, \ldots, S_n$ be discrete one-dimensional sample spaces. then the n-dimensional sample space might be

$$S = S_1 \times S_2 \times \cdots \times S_n.$$

We write the n-dimensional random variable as $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$.

First consider a two-dimensional discrete random variable

$$S = S_1 \times S_2, \quad \boldsymbol{X} = (X_1, X_2).$$

The *joint pmf* for (X_1, X_2) is denoted by

$$p_{X_1,X_2}(x_1, x_2) = \mathrm{P}_{X_1,X_2}(\{x_1, x_2\}) = \mathrm{P}_{X_1,X_2}(X_1 = x_1,\ X_2 = x_2).$$

Example 2.9.1. Toss two dice and assume equally likely points in S.

$$S_1 = \{1,2,3,4,5,6\},\ S_2 = \{1,2,3,4,5,6\},\ \text{and } S = S_1 \times S_2 :$$

Thus,

$$p_{X_1,X_2}(2,3) = \mathrm{P}_{X_1,X_2}(\{2,3\}) = \mathrm{P}_{X_1,X_2}(X_1 = 2,\ X_2 = 3) = \frac{1}{36},$$
$$p_{X_1,X_2}(0,3) = \mathrm{P}_{X_1,X_2}(\{0,3\}) = \mathrm{P}_{X_1,X_2}(X_1 = 0,\ X_2 = 3) = 0. \qquad \square$$

Definition 2.9.1 (Marginal pmf). For the two-dimensional case, the *marginal probability mass function* for X_1 s given by

$$p_{X_1}(x_1) = \sum_{\text{all } x_2} p_{X_1,X_2}(x_1, x_2).$$

Likewise, the marginal pmf for X_2 is given by

$$p_{X_2}(x_2) = \sum_{\text{all } x_1} p_{X_1,X_2}(x_1, x_2).$$

Definition 2.9.2 (Joint Distribution Function). For the two-dimensional case, the *joint distribution function* s given by

$$F_{X_1,X_2}(x_1, x_2) = \sum_{y_2 \le x_2} \sum_{y_1 \le x_1} p_{X_1,X_2}(x_1, x_2).$$

Example 2.9.2. Let $S_1 = \{1, 2, 3, 4, 5\}$, $S_2 = \{18, 19, 20, 21\}$, $S = S_1 \times S_2$, and let the joint and marginal probabilities be given by

$x_1 \setminus x_2$	18	19	20	21	$p_{X_1}(x_1)$
1	0.00	0.02	0.05	0.00	0.07
2	0.01	0.02	0.06	0.00	0.09
3	0.07	0.14	0.18	0.06	0.45
4	0.06	0.08	0.09	0.02	0.25
5	0.03	0.07	0.03	0.01	0.14
$p_{X_2}(x_2)$	0.17	0.33	0.41	0.09	1.00

Thus,

$$p_{X_1,X_2}(4, 19) = \mathrm{P}_{X_1,X_2}(X_1 = 4,\ X_2 = 19) = 0.08,$$

and for marginal pmf's,

$$\begin{aligned} p_{X_1}(2) &= \sum_{x_2=18}^{21} p_{X_1,X_2}(2, x_2) \\ &= p_{X_1,X_2}(2, 18) + p_{X_1,X_2}(2, 19) + p_{X_1,X_2}(2, 20) + p_{X_1,X_2}(2, 21) \\ &= 0.01 + 0.02 + 0.06 + 0.00 && \text{from the table} \\ &= 0.09. \end{aligned}$$

$$\begin{aligned} p_{X_1}(3) &= \sum_{x_2=18}^{21} p_{X_1,X_2}(3, x_2) \\ &= p_{X_1,X_2}(3, 18) + p_{X_1,X_2}(3, 19) + p_{X_1,X_2}(3, 20) + p_{X_1,X_2}(3, 21) \\ &= 0.07 + 0.14 + 0.18 + 0.06 && \text{from the table} \\ &= 0.45. \end{aligned}$$

$$\begin{aligned} p_{X_2}(19) &= \sum_{x_1=1}^{5} p_{X_1,X_2}(x_1, 19) \\ &= p_{X_1,X_2}(1, 19) + p_{X_1,X_2}(2, 19) + p_{X_1,X_2}(3, 19) + p_{X_1,X_2}(4, 19) + p_{X_1,X_2}(5, 19) \\ &= 0.02 + 0.02 + 0.14 + 0.08 + 0.07 && \text{from the table} \\ &= 0.33. \end{aligned}$$

$$\begin{aligned} p_{X_2}(20) &= \sum_{x_1=1}^{5} p_{X_1,X_2}(x_1, 20) \\ &= p_{X_1,X_2}(1, 20) + p_{X_1,X_2}(2, 20) + p_{X_1,X_2}(3, 20) + p_{X_1,X_2}(4, 20) + p_{X_1,X_2}(5, 20) \\ &= 0.05 + 0.06 + 0.18 + 0.09 + 0.03 && \text{from the table} \\ &= 0.41. \end{aligned}$$

$x_1 \setminus x_2$	18	19	20	21	$p_{X_1}(x_1)$
1	0.00	0.02	0.05	0.00	0.07
2	0.01	0.02	0.06	0.00	0.09
3	0.07	0.14	0.18	0.06	0.45
4	0.06	0.08	0.09	0.02	0.25
5	0.03	0.07	0.03	0.01	0.14
$p_{X_2}(x_2)$	0.17	0.33	0.41	0.09	1.00

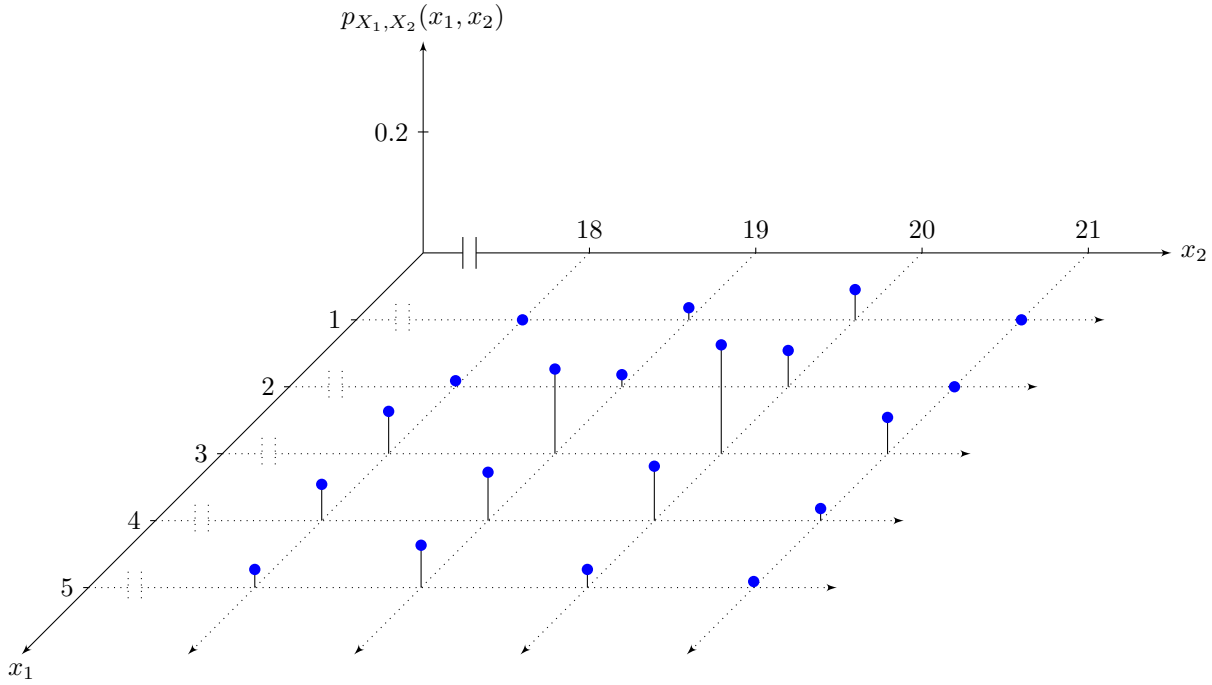
For the joint distribution function,

$$\begin{aligned} F_{X_1,X_2}(2,19) &= \sum_{x_2=18}^{19} \sum_{x_1=1}^{2} p_{X_1,X_2}(x_1,x_2) \\ &= p_{X_1,X_2}(1,18) + p_{X_1,X_2}(2,18) + p_{X_1,X_2}(1,19) + p_{X_1,X_2}(2,19) \\ &= 0.00 + 0.01 + 0.02 + 0.02 \\ &= 0.05. \end{aligned}$$

Furthermore, note that

$$\begin{aligned} \mathrm{P}_{X_1,X_2}(X_1 = 3 \text{ and } 18 \leq X_2 \leq 21) &= p_{X_1}(3) = 0.45, \\ \mathrm{P}_{X_1,X_2}(X_1 = 3 \text{ and } 18 < X_2 < 21) &= p_{X_1,X_2}(3,19) + p_{X_1,X_2}(3,20) = 0.32. \end{aligned}$$

A graph of our joint probabilities looks like this:



□

Example 2.9.3 (Drawing from an urn without replacement). Suppose we draw two balls without replacement from an urn containing 10 balls:

$$\left\{ \begin{array}{c} \text{5 red balls: \# 1, 2, 3, 4, 5} \\ \text{5 white balls: \# 6, 7, 8, 9, 10} \end{array} \right\}$$

Assume equally likely points in the sample space:

$$S = \{(x_1, x_2) : \ x_1 = 1, 2, \ldots, 10; \ x_2 = 1, 2, \ldots, 10; \ x_1 \neq x_2\}.$$

Let the random variable X_i denote the outcome of the i^{th} draw:

$$p_{X_1,X_2}(x_1, x_2) = \frac{1}{10 \cdot 9} = \frac{1}{90} \quad \text{for } (x_1, x_2) \in S.$$

$x_1 \setminus x_2$	1	2	3	$\cdots$	10	$p_{X_2}(x_2)$
1	0	$\frac{1}{90}$	$\frac{1}{90}$	$\cdots$	$\frac{1}{90}$	$\frac{9}{90}$
2	$\frac{1}{90}$	0	$\frac{1}{90}$	$\cdots$	$\frac{1}{90}$	$\frac{9}{90}$
3	$\frac{1}{90}$	$\frac{1}{90}$	0	$\cdots$	$\frac{1}{90}$	$\frac{9}{90}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	0	$\vdots$	$\vdots$
10	$\frac{1}{90}$	$\frac{1}{90}$	$\frac{1}{90}$	$\cdots$	0	$\frac{9}{90}$
$p_{X_1}(x_1)$	$\frac{9}{90}$	$\frac{9}{90}$	$\frac{9}{90}$	$\cdots$	$\frac{9}{90}$	$\frac{9}{90}$

As shown in the above table, our marginal pmf's for X_1 and X_2 are

$$p_{X_1}(x_1) = \sum_{x_2=1}^{10} p_{X_1,X_2}(x_1, x_2) = \frac{9}{90} = \frac{1}{10} \quad \text{for } x_1 \in \{1, 2, \ldots, 10\},$$

$$p_{X_2}(x_2) = \sum_{x_1=1}^{10} p_{X_1,X_2}(x_1, x_2) = \frac{9}{90} = \frac{1}{10} \quad \text{for } x_2 \in \{1, 2, \ldots, 10\}.$$

A couple points to ponder:

- We know from above that $p_{X_2}(5) = \mathrm{P}(X_2 = 5) = \frac{9}{90}$. Another way to think about it is that

$$\begin{aligned} \mathrm{P}(X_2 = 5) &= \mathrm{P}(\{(x_1, 5) : x_1 = 1, \ldots 10; \ x_1 \neq 5\}) \\ &= \sum_{\substack{x_1=1 \\ x_1 \neq 5}}^{10} \mathrm{P}(\{(x_1, 5)\}) \\ &= \frac{9}{90}. \end{aligned}$$

- Note that $p_{X_1,X_2}(x_1, x_2) \neq P_{X_1}(x_1) p_{X_2}(x_2)$.

□

For the n-dimensional case, the joint pmf is defined by

$$p_{X_1,X_2,\ldots,X_n}(x_1,\ldots,x_n) = p_{\boldsymbol{X}}(x_1,\ldots,x_n) = \mathrm{P}_{\boldsymbol{X}}(\{(x_1,\ldots,x_n)\}).$$

The joint distribution function is defined as

$$F_{\boldsymbol{X}}(x_1,\ldots,x_n) = \sum_{y_n \le x_n} \sum_{y_{n-1} \le x_{n-1}} \cdots \sum_{y_1 \le x_1} p_{\boldsymbol{X}}(x_1,\ldots,x_n).$$

Example 2.9.4. Suppose we toss three unbiased dice:

$$\begin{aligned} S &= S_1 \times S_2 \times S_3 && \text{where } S_i = \{1,2,3,4,5,6\} \\ &= \{(x_1,x_2,x_3) :\ x_i = 1,2,3,4,5,6\}. \end{aligned}$$

We define

$$p_{X_1,X_2,X_3}(x_1,x_2,x_3) = \frac{1}{6\cdot 6\cdot 6} = \frac{1}{216} \quad \forall (x_1,x_2,x_3) \in S.$$

Thus,

$$\begin{aligned} F_{\boldsymbol{X}}(1,2,4) &= \sum_{x_3=1}^{4}\sum_{x_2=1}^{2}\sum_{x_1=1}^{1} p_{\boldsymbol{X}}(x_1,x_2,x_3) \\ &= \sum_{x_3=1}^{4}\sum_{x_2=1}^{2}\sum_{x_1=1}^{1} \frac{1}{216} \\ &= \frac{4\cdot 2\cdot 1}{216} \\ &= \frac{8}{216} \\ &\approx 3.70\%. \end{aligned}$$

□

2.10 n-Dimensional Continuous Random Variables

In this case, the sample space is $\mathbb{R}^n = \mathbb{R}\times\mathbb{R}\times\cdots\times\mathbb{R}$. The random variable is denoted by $(x_1,\ldots,x_n)$. The joint probability density function is denoted by $f_{\boldsymbol{X}}(x_1,\ldots,x_n)$, and the joint distribution function is given by

$$\begin{aligned} F_{\boldsymbol{X}}(x_1,\ldots,x_n) &= \int_{-\infty}^{x_n}\int_{-\infty}^{x_{n-1}}\cdots\int_{-\infty}^{x_1} f_{\boldsymbol{X}}(y_1,\ldots,y_n)\ dy_1\cdots dy_{n-1}dy_n \\ &= \mathrm{P}_{\boldsymbol{X}}(X_1 \le x_1,\ X_2 \le x_2,\ \ldots,\ X_n \le x_n). \end{aligned}$$

For the two-dimensional case, the marginal densities are

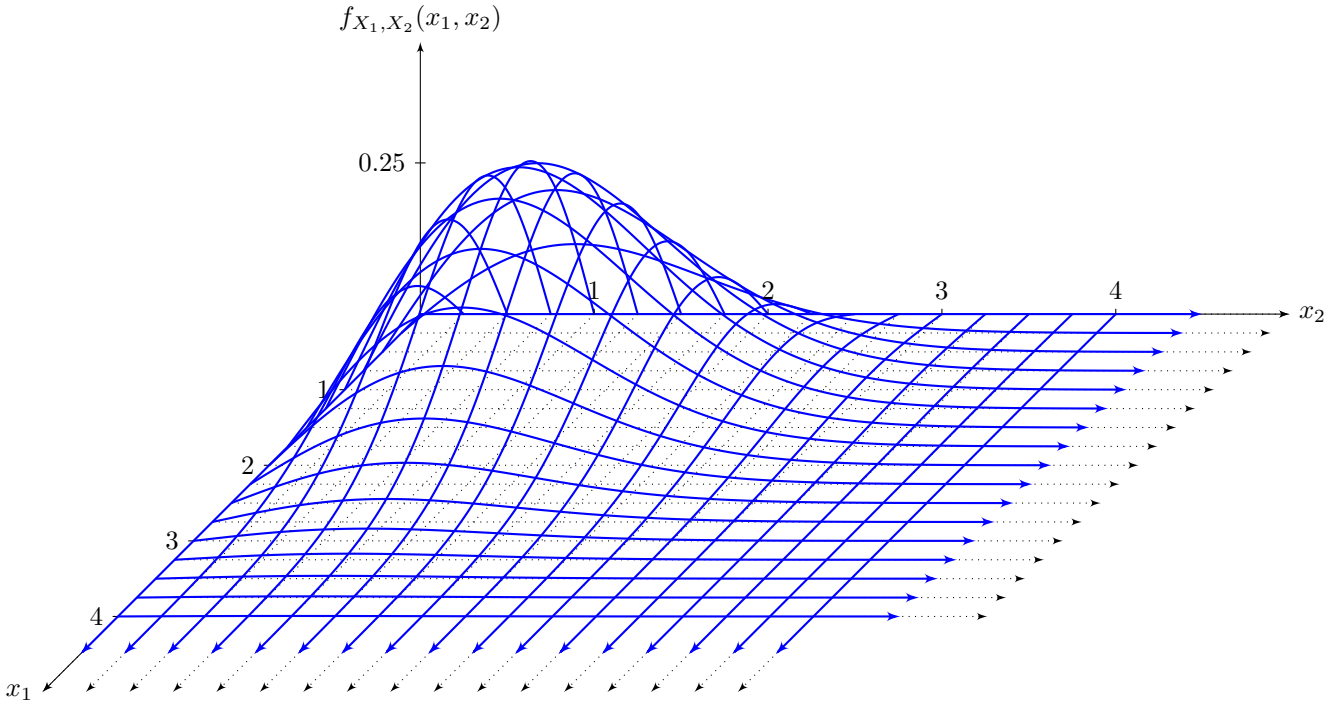
$$\begin{aligned} f_{X_1}(x_1) &= \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)\ dx_2, \\ f_{X_2}(x_2) &= \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)\ dx_1. \end{aligned}$$

Example 2.10.1. Suppose at two points in a room, one measures the intensity of sound caused by general background noise. Let X_1 and X_2 be the r.v.'s representing the intensity of sound at two points:

X_1 is the intensity of sound measured at a point on the floor,
X_2 is the intensity of sound measured at a point on the ceiling.

Assume that the joint pdf for X_1 and X_2 is given by

$$f_{X_1,X_2}(x_1,x_2) = x_1 x_2 \exp\left(-\frac{x_1^2+x_2^2}{2}\right), \quad x_1, x_2 > 0.$$



Then the marginal distributions for X_1 and X_2 are given by

$$\begin{aligned} f_{X_1}(x_1) &= \int_0^\infty x_1 x_2 \exp\left(-\frac{x_1^2+x_2^2}{2}\right) dx_2 \\ &= x_1\left[-\exp\left(-\frac{x_1^2+x_2^2}{2}\right)\right]_{x_2=0}^{x_2=\infty} \\ &= x_1 e^{-x_1^2/2}, \quad x_1 > 0. \end{aligned}$$

$$f_{X_2}(x_2) = x_2 e^{-x_2^2/2}, \quad x_2 > 0. \qquad \text{by symmetry}$$

Note that in this case, $f_{\boldsymbol{X}}(x_1,x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$.

The distribution function for $x_1, x_2 > 0$ is

$$\begin{aligned}
F_{\boldsymbol{X}}(x_1, x_2) &= \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f_{\boldsymbol{X}}(t_1, t_2)\ dt_1 dt_2 \\
&= \int_0^{x_2} \int_0^{x_1} t_1 t_2 \exp\left(-\frac{t_1^2 + t_2^2}{2}\right)\ dt_1 dt_2 \\
&= \int_0^{x_2} \left[\int_0^{x_1} t_1 e^{-t_1^2/2}\ dt_1\right]\ t_2 e^{-t_2^2/2} dt_2 && \text{since } t_2 e^{-t_2^2/2} \text{ is a constant with respect to } t_1 \\
&= \int_0^{x_1} t_1 e^{-t_1^2/2}\ dt_1 \int_0^{x_2} t_2 e^{-t_2^2/2}\ dt_2 && \text{since } \int_0^{x_1} t_1 e^{-t_1^2/2}\ dt_1 \text{ is a constant with respect to } t_2 \\
&= \left(1 - e^{-x_1^2/2}\right)\left(1 - e^{-x_2^2/2}\right)
\end{aligned}$$

For example,

$$\begin{aligned}
F_{\boldsymbol{X}}(1,1) &= \mathrm{P}(X_1 \leq 1,\ X_2 \leq 1) \\
&= \left(1 - e^{-1/2}\right)\left(1 - e^{-1/2}\right) \\
&\approx 15.48\%.
\end{aligned}$$

□

3 Dependent and Independent Events

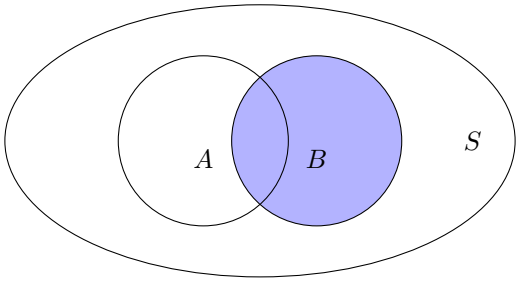
3.1 Conditional Probability

We want to know the probability of some event A given (having the information) that some event B has occurred.

Definition 3.1.1. If $\mathrm{P}(B) > 0$, then then *conditional probability of A given B*, denoted $\mathrm{P}(A|B)$, is given by

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

In words, $\mathrm{P}(A|B)$ represents the re-evaluation of the probability of A in light of the information that B has occurred. It reflects the fact that our sample space has shrunk from S to B.

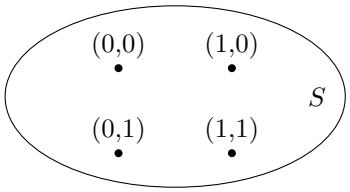


Example 3.1.1. Consider a family with two children. Assume that each child is equally likely to be a boy or a girl.

Let A be the event that the first child is a boy,
Let B be the event that the second child is a boy.

Then $A \cup B$ is the event that at least one child is a boy, and $A \cap B$ is the event that both children are boys. Let the r.v. X and Y denote the sex of the first and second child, respectively. Then our sample space becomes

$$S = \{(x, y) : \ x = 0, 1; \ y = 0, 1\}:$$



$\mathrm{P}(A) = \mathrm{P}(B) = \frac{1}{2}$, $\mathrm{P}(AB) = \frac{1}{4}$ and $\mathrm{P}(A \cup B) = \frac{3}{4}$, so

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(AB)}{\mathrm{P}(B)} = \frac{1/4}{1/2} = \frac{1}{2},$$

$$\mathrm{P}(AB|B) = \frac{\mathrm{P}(ABB)}{\mathrm{P}(B)} = \frac{\mathrm{P}(AB)}{\mathrm{P}(B)} = \frac{1/4}{1/2} = \frac{1}{2},$$

$$\mathrm{P}(AB|A \cup B) = \frac{\mathrm{P}(AB \cap (A \cup B))}{\mathrm{P}(A \cup B)} = \frac{\mathrm{P}(AB)}{\mathrm{P}(A \cup B)} = \frac{1/4}{3/4} = \frac{1}{3}.$$

Note that $\mathrm{P}(A|B) = \mathrm{P}(A)$ and that

$$\mathrm{P}(AB) = \mathrm{P}(A|B)\mathrm{P}(B) = \mathrm{P}(B|A)\mathrm{P}(A). \tag{3.1.1}$$

□

In fact, Equation (3.1.1) can be generalized.

Theorem 3.1.1 (Multiplication Rule). Let $A_1, A_2, \ldots, A_n$ be defined events on the sample space S. Then

$$\mathrm{P}(A_1A_2A_3\cdots A_n) = \mathrm{P}(A_1)\mathrm{P}(A_2|A_1)\mathrm{P}(A_3|A_1A_2)\cdots\mathrm{P}(A_n|A_1A_2\cdots A_{n-1}). \tag{3.1.2}$$

This is called *the multiplication rule*, and using it to solve a problem is called *conditioning*.

Proof: We simply apply the definition of conditional probability multiple times:

$$\begin{aligned}\mathrm{P}(A_1A_2A_3\cdots A_n) &= \mathrm{P}(A_1)\cdot\frac{\mathrm{P}(A_1A_2)}{\mathrm{P}(A_1)}\cdot\frac{\mathrm{P}(A_1A_2A_3)}{\mathrm{P}(A_1A_2)}\cdots\frac{\mathrm{P}(A_1A_2\cdots A_n)}{\mathrm{P}(A_1A_2\cdots A_{n-1})}\\ &= \mathrm{P}(A_1)\mathrm{P}(A_2|A_1)\mathrm{P}(A_3|A_1A_2)\cdots\mathrm{P}(A_n|A_1A_2\cdots A_{n-1}).\end{aligned}$$

■

Theorem 3.1.2. Let S be a sample space and $\mathrm{P}(\cdot)$ be the probability function defined over the set of events Ψ. Then for any $B \in \Psi$, if $\mathrm{P}(B) > 0$, then $\mathrm{P}(\cdot|B)$ is a probability function. That is,

(a) $\mathrm{P}(A|B) \geq 0 \quad \forall A \in \Psi$,

(b) $\mathrm{P}(S|B) = 1$, and

(c) If $A_1, A_2, \ldots, A_n$ are disjoint sets (events), then $\mathrm{P}\left(\bigcup_{k=1}^{n} A_k \middle| B\right) = \sum_{k=1}^{n}\mathrm{P}(A_k|B)$.

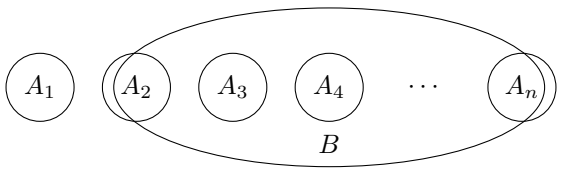
Proof:

(a) $\mathrm{P}(A|B) = \dfrac{\mathrm{P}(AB)}{\mathrm{P}(B)} \geq 0$, since $\mathrm{P}(\cdot)$ is a probability function.

(b) $\mathrm{P}(S|B) = \dfrac{\mathrm{P}(SB)}{\mathrm{P}(B)} = \dfrac{\mathrm{P}(B)}{\mathrm{P}(B)} = 1.$

(c) $\mathrm{P}\left(\bigcup_{k=1}^{n} A_k \middle| B\right) = \dfrac{\mathrm{P}\left([\cup_k A_k] \cap B\right)}{\mathrm{P}(B)} = \dfrac{\mathrm{P}\left(\cup_k A_kB\right)}{\mathrm{P}(B)} \underbrace{=}_{*} \sum_{k=1}^{n}\dfrac{\mathrm{P}(A_kB)}{\mathrm{P}(B)} = \sum_{k=1}^{n}\mathrm{P}(A_k|B),$

where step $*$ is true because the sets $\{A_kB\}_k$ are disjoint.



■

Since $\mathrm{P}(\cdot|B)$ is a probability function (as proven above), every theorem we have proven for a probability function holds for a conditional probability function as well.

Definition 3.1.2. A *partition* of a set S is a collection of disjoint nonempty sets $\{S_i\}$ whose union is equal to S. That is,

- $S_i \neq \varnothing \ \ \forall i$.
- $S_i \cap S_j = \varnothing$ for $i \neq j$.
- $\bigcup_i S_i = S$.

Note that there need not be a finite number of sets in in the collection $\{S_i\}$.

Example 3.1.2. (a) Let $S = \{1, 2, 3, 4, 5\}$. A possible partition of S is

$$\{1\} \ \ \{2\} \ \ \{3\} \ \ \{4\} \ \ \{5\},$$

while another is

$$\{1, 2, 3, 4\} \ \ \{5\},$$

and still another is

$$\{1, 2, 3, 4, 5\}.$$

(b) Let $S = \mathbb{Z}^+$, the set of natural numbers. A possible partition of S is

$$\{1\} \ \ \{2\} \ \ \{3\} \ \ \{4\} \ \ \{5\} \ \ \ldots,$$

while another is

$$\{1, 2, 3, 4\} \ \ \{5, 6, 7, \ldots\},$$

and still another is

$$\{1, 2, 3, 4, 5, 6, 7, \ldots\}.$$ □

Theorem 3.1.3 (Theorem of Total Probabilities). Let S be a minimal sample space and let $\{S_i\}$ be a partition of S. Then for every event $A \subset S$,

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

Proof:

$$P(A) = \mathrm{P}(A \cap S) = \mathrm{P}(A \cap (\cup_i S_i)) = \mathrm{P}(\cup_i A S_i) = \sum_{i=1}^{n} \mathrm{P}(AS_i) = \sum_{i=1}^{n} \frac{\mathrm{P}(AS_i)\mathrm{P}(S_i)}{\mathrm{P}(S_i)} = \sum_{i=1}^{n} \mathrm{P}(A|S_i)\mathrm{P}(S_i).$$

Note that in the penultimate equality, we must have that $\mathrm{P}(S_i) > 0 \ \forall i$. How do we know that this is true? ■

Example 3.1.3. A box has n_1 tags numbered 1 and n_2 tags numbered 2. A tag is selected at random (each tag is equally likely). We have two urns:

$$\text{urn \#1 contains } r_1 \text{ red and } b_1 \text{ black balls,}$$
$$\text{urn \#2 contains } r_2 \text{ red and } b_2 \text{ black balls.}$$

If tag $\#i$ is selected, one goes to urn $\#i$ and selects a ball at random. Thus, our sample space is

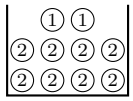
$$S = \{(x, y) : \ x = 1, 2; \ y = 1, \ldots, r_1 + b_1 \text{ if } x = 1; \ y = 1, \ldots, r_2 + b_2 \text{ if } x = 2\}.$$

Let R be the event that the ball drawn is red, and H_i be the event that tag $\#i$ is drawn. What is $\mathrm{P}(R)$? Using the theorem of total probabilities,

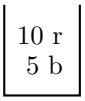
$$\mathrm{P}(R) = \mathrm{P}(R|H_1)\mathrm{P}(H_1) + \mathrm{P}(R|H_2)\mathrm{P}(H_2) = \left(\frac{r_1}{r_1 + b_1}\right)\left(\frac{n_1}{n_1 + n_2}\right) \ + \ \left(\frac{r_2}{r_2 + b_2}\right)\left(\frac{n_2}{n_1 + n_2}\right).$$

For instance, suppose $r_1 = 10$, $b_1 = 5$, $r_2 = 5$, $b_2 = 10$.

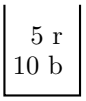
- Let $n_1 = 2$ and $n_2 = 8$:



box



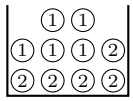
urn 1



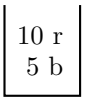
urn 2

$$\Rightarrow \qquad \mathrm{P}(R) = \left(\tfrac{10}{15}\right)\left(\tfrac{2}{10}\right) \ + \ \left(\tfrac{5}{15}\right)\left(\tfrac{8}{10}\right) = \tfrac{6}{15} < \tfrac{1}{2}.$$

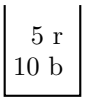
- Let $n_1 = 5$ and $n_2 = 5$:



box



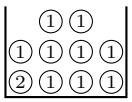
urn 1



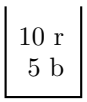
urn 2

$$\Rightarrow \qquad \mathrm{P}(R) = \left(\tfrac{10}{15}\right)\left(\tfrac{5}{10}\right) \ + \ \left(\tfrac{5}{15}\right)\left(\tfrac{5}{10}\right) = \tfrac{1}{2}.$$

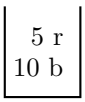
- Let $n_1 = 9$ and $n_2 = 1$:



box



urn 1



urn 2

$$\Rightarrow \qquad \mathrm{P}(R) = \left(\tfrac{10}{15}\right)\left(\tfrac{9}{10}\right) \ + \ \left(\tfrac{5}{15}\right)\left(\tfrac{1}{10}\right) = \tfrac{95}{150} > \tfrac{1}{2}.$$

□

Theorem 3.1.4 (Bayes' Rule). Let S be a minimal sample space and let $\{H_i\}$ be a partition of S. Then for every event $A \subset S$ and every i,

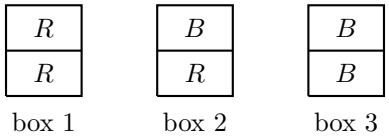
$$\mathrm{P}(H_i|A) = \frac{\mathrm{P}(A|H_i)\mathrm{P}(H_i)}{\sum_j \mathrm{P}(A|H_j)\mathrm{P}(H_j)}.$$

Proof: Using the definition of conditional probability and the theorem of total probability,

$$\mathrm{P}(H_i|A) = \frac{\mathrm{P}(A \cap H_i)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A|H_i)\mathrm{P}(H_i)}{\sum_j \mathrm{P}(A|H_j)\mathrm{P}(H_j)}.$$

■

Example 3.1.4. Suppose we have equally likely boxes, and in each box we have equally likely drawers. In each drawer there is one ball, colored red (R) or blue (B):



Let H_i be the event that the i^{th} box is chosen. Suppose we randomly choose a drawer from a randomly chosen box. Given that a red ball was found in our drawer, what is the probability that the first box was chosen? Intuitively, we might guess the answer to be $\frac{2}{3}$. To prove this, we can use Bayes' rule:

$$\mathrm{P}(H_1|R) = \frac{\mathrm{P}(R|H_1)\mathrm{P}(H_1)}{\sum_{j=1}^{3}\mathrm{P}(R|H_j)\mathrm{P}(H_j)} = \frac{1\cdot\frac{1}{3}}{\left(1\cdot\frac{1}{3}\right) + \left(0\cdot\frac{1}{3}\right) + \left(\frac{1}{2}\cdot\frac{1}{3}\right)} = \frac{2}{3}. \qquad \square$$

Example 3.1.5. Suppose a factory has two machines, A and B, which make 60% and 40% of the total population of parts, respectively. Of their output, machine A produces 3% defective items, while B produces 5% defective items. Find the probability that a given defective part was produced by machine B.

Let D be the event that an item is defective. We want to find

$$\mathrm{P}(B|D) = \frac{\mathrm{P}(D|B)\mathrm{P}(B)}{\mathrm{P}(D|A)\mathrm{P}(A) + \mathrm{P}(D|B)\mathrm{P}(B)} = \frac{(0.05)(0.40)}{(0.03)(0.60) + (0.05)(0.40)} \approx 52.63\%. \qquad \square$$

Example 3.1.6 (Cancer Diagnosis). Let A denote the event that a person tested has cancer, and C denote the event that a test states that a person has cancer. This diagnostic test has the following properties:

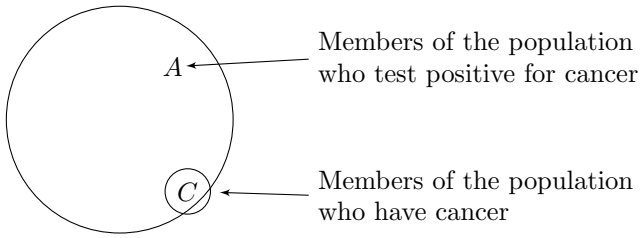
$$\begin{aligned} \mathrm{P}(A|C) &= 0.95, & \mathrm{P}(C) &= 0.005, \\ \mathrm{P}(A^c|C^c) &= 0.95, & \mathrm{P}(C^c) &= 1 - 0.005 = 0.995, \\ \mathrm{P}(A|C^c) &= 1 - 0.95 = 0.05. \end{aligned}$$

We want to find the probability that a person who is diagnosed with cancer actually has cancer:

$$\mathrm{P}(C|A) = \frac{\mathrm{P}(A|C)\mathrm{P}(C)}{\mathrm{P}(A|C)\mathrm{P}(C) + \mathrm{P}(A|C^c)\mathrm{P}(C^c)} = \frac{(0.95)(0.005)}{(0.95)(0.005) + (0.05)(0.995)} \approx 8.72\%.$$

Overall, the test will detect cancer in 95% of the cases where cancer is present. But in only 8.72% of the cases where the test is positive is cancer actually present. Care must be taken in interpreting this result – what we actually mean is that *if you take a random person off the street and test her for cancer and you get a positive result, there is an 8.72% chance that she actually has cancer.* This takes into account the whole population, the vast majority of which does not have cancer! This is very different from testing someone who shows some symptoms of possibly having cancer (which is *not* representative of the entire population).

$$\begin{aligned} \mathrm{P}(C|A) &= 8.72\%, \\ \mathrm{P}(C^c|A) &= 91.28\%. \end{aligned}$$



$\square$

3.2 Independence

Definition 3.2.1 (Independence). Let C be a collection of events. That is, $C \subset \Psi$ for some sample space S. The events in C are said to be *independent* iff the probability of the joint occurrence of any finite number of them equals the product of their probabilities. That is,

$$\mathrm{P}(A_1 A_2 \cdots A_n) = \mathrm{P}(A_1)\mathrm{P}(A_2)\cdots\mathrm{P}(A_n) \quad \text{for any } A_1, \ldots, A_n \in C. \tag{3.2.3}$$

If C consists of two events A and B, then these two events are independent iff

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

If C consists of three events A, B and C, then these three events are independent iff

$$\mathrm{P}(AB) = \mathrm{P}(A)\mathrm{P}(B), \quad \mathrm{P}(AC) = \mathrm{P}(A)\mathrm{P}(C), \quad \mathrm{P}(BC) = \mathrm{P}(B)\mathrm{P}(C), \quad \text{and} \quad \mathrm{P}(ABC) = \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C).$$

If C has an infinite number of events, than (3.2.3) must hold for any finite collection of events in C.

Example 3.2.1. Suppose we draw with replacement a sample size of 2 from an urn containing 4 white balls and 2 red balls.

Let A be the event that the first ball drawn is white,
Let B be the event that the second ball drawn is white.

Thus,

$$\begin{aligned}
S &= \{(x,y):\ x = 1,\ldots,6;\ y = 1,\ldots 6\} &\Rightarrow&\quad \text{Size of } S \text{ is } 6\cdot 6 = 36,\\
A &= \{(x,y):\ x = 1,\ldots,4;\ y = 1,\ldots 6\} &\Rightarrow&\quad \text{Size of } A \text{ is } 4\cdot 6 = 24,\\
B &= \{(x,y):\ x = 1,\ldots,6;\ y = 1,\ldots 4\} &\Rightarrow&\quad \text{Size of } B \text{ is } 6\cdot 4 = 24,\\
AB &= \{(x,y):\ x = 1,\ldots,4;\ y = 1,\ldots 4\} &\Rightarrow&\quad \text{Size of } AB \text{ is } 4\cdot 4 = 16,
\end{aligned}$$

Then assuming equally lilkely points in S,

$$\mathrm{P}(A) = \frac{24}{36} = \frac{2}{3}, \quad \mathrm{P}(B) = \frac{24}{36} = \frac{2}{3}, \quad \mathrm{P}(AB) = \frac{16}{36} = \frac{4}{9}.$$

Note that $\mathrm{P}(AB) = \frac{4}{9} = \left(\frac{2}{3}\right)\left(\frac{2}{3}\right) = \mathrm{P}(A)\mathrm{P}(B)$. Then we can conclude that A and B are independent. □

At this point, one might wonder if all of (3.2.3) really is necessary for independence. Would it be enough simply to have *pairwise independence*? That is, if we know that

$$\mathrm{P}(A_i A_j) = \mathrm{P}(A_i)\mathrm{P}(A_j) \quad \forall i, j \in \{1, 2, \ldots, n\}, \tag{3.2.4}$$

can we then assume (3.2.3)? The answer is no.

Example 3.2.2 (Counterexample). This counterexample is due to Russian probabilist Sergei Bernstein (1880-1968). Suppose S consists of 4 equally likely points x_1, x_2, x_3 and x_4. Let

$$A = \{x_1, x_2\}, \quad B = \{x_1, x_3\}, \quad C = \{x_1, x_4\} \quad \Rightarrow \quad \mathrm{P}(A) = \mathrm{P}(B) = \mathrm{P}(C) = \frac{1}{2}.$$

Also, $AB = BC = AC = \{x_1\}$, so that $\mathrm{P}(AB) = \mathrm{P}(BC) = \mathrm{P}(AC) = \frac{1}{4} = \mathrm{P}(A)\mathrm{P}(B) = \mathrm{P}(B)\mathrm{P}(C) = \mathrm{P}(A)\mathrm{P}(C)$. However,

$$\mathrm{P}(ABC) = \mathrm{P}(\{x_1\}) = \frac{1}{4} \neq \frac{1}{8} = \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C).$$

Therefore (3.2.4) is *not* equivalent to (3.2.3). □

Recall that $\mathrm{P}(AB) = \mathrm{P}(A|B)\mathrm{P}(B) = \mathrm{P}(B|A)\mathrm{P}(A)$. This is true for *any* two events A and B. From above, we now know that A and B are independent iff $\mathrm{P}(AB) = \mathrm{P}(A)\mathrm{P}(B)$. From these two statements, we might deduce that A and B are independent iff $\mathrm{P}(A) = \mathrm{P}(A|B)$, or iff $\mathrm{P}(B) = \mathrm{P}(B|A)$. We shall prove this below.

Theorem 3.2.1. If $\mathrm{P}(B) > 0$, then a necessary and sufficient condition that the events A and B are independent is that

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

Proof:

($\Rightarrow$) Assume independence. Then $\mathrm{P}(AB) = \mathrm{P}(A)\mathrm{P}(B)$, so that

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

($\Leftarrow$) Now assume that $\mathrm{P}(A|B) = \mathrm{P}(A)$. Then

$$\mathrm{P}(AB) = \mathrm{P}(A|B)\mathrm{P}(B) = \mathrm{P}(A)\mathrm{P}(B),$$

which shows independence.

■

Example 3.2.3. Toss a fair coin 5 times and assume that the outcome on any one toss is independent of the outcomes of all other tosses. What is the probability of getting five heads?

Let H_i be the event that a head is obtained on the i^{th} toss. Our sample space is

$$S = \{(x_1, \ldots, x_5) : \ x_i = H, T; \ i = 1, \ldots, 5\} \quad \Rightarrow \quad \text{Size of } S \text{ is } 2^5.$$

Assume all points in S are equally likely. Thus, the probability of any one point in S is $\frac{1}{2^5} = \frac{1}{32}$. In particular, the probability of getting our five heads (H, H, H, H, H) is $\frac{1}{32}$. However, we can also get that answer a different way. Note that

$$H_i = \{(x_1, \ldots, x_5) : \ x_i = H, \ x_{j \neq i} = H, T\} \quad \Rightarrow \quad \text{Size of } H_i \text{ is } 2^4.$$

Therefore, $\mathrm{P}(H_i) = \frac{2^4}{2^5} = \frac{1}{2}\ \forall i$, so that

$$\mathrm{P}(\{(H,H,H,H,H)\}) = \mathrm{P}(H_1H_2H_3H_4H_5) = \mathrm{P}(H_1)\mathrm{P}(H_2)\mathrm{P}(H_3)\mathrm{P}(H_4)\mathrm{P}(H_5) = \left(\frac{1}{2}\right)^5 = \frac{1}{2^5} = \frac{1}{32},$$

as we had before. □

The purpose of doing the above problem two different ways is to show just that – that there is sometimes (often!) two or more ways to find the answer to a given problem. While this example is trivial in that both ways are very simple, there are often times where one way to get an answer is (much) easier than another way. Therefore, if finding a solution to a problem is very difficult, it might be useful to stop and think of another approach to the problem.

Definition 3.2.2 (Independence of Random Variables). Random variables $X_1, X_2, \ldots, X_n$ are *independent* iff their joint pmf or pdf can be factored into the product of their individual pmf's or pdf's. That is, iff

$$p_{\boldsymbol{X}}(x_1, x_2, \ldots, x_n) = p_{X_1}(x_1)p_{X_2}(x_2)\cdots p_{X_n}(x_n) \qquad \forall (x_1, x_2, \ldots, x_n) \in \overline{\mathbb{R}}^n \tag{3.2.5}$$

if $X_1, X_2, \ldots, X_n$ are discrete, and

$$f_{\boldsymbol{X}}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n) \qquad \forall (x_1, x_2, \ldots, x_n) \in \overline{\mathbb{R}}^n \tag{3.2.6}$$

if $X_1, X_2, \ldots, X_n$ are continuous. Furthermore, the set of random variables in an infinite sequence of them $X_1, X_s, \ldots$ are independent if every finite subset of them is independent, as described above.

It may be tempting to ask why the above definition of independent *random variables* is slightly different from the definition of independent *events* as stated in (3.2.3). That is, n events are independent only if every set of k of them are mutually independent *for every value of* $k \leq n$, whereas for random variables it suffices to just check if all n of them are independent. Why is this so?

The answer lies in the fact that (3.2.3) must be true for *one* given set of events $A_1, \ldots, A_n$, whereas (3.2.5) and (3.2.6) must hold for *all* values of $x_1, \ldots, x_n$. Specifically, it can be shown (e.g., Parzen (1960, pp. 294-295)) that (3.2.5) and (3.2.6) are both equivalent to the condition that

$$F_{\boldsymbol{X}}(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n) \qquad \forall (x_1, x_2, \ldots, x_n) \in \overline{\mathbb{R}}^n,$$

which is equivalent to

$$\mathrm{P}(X_1 \leq x_1,\ X_2 \leq x_2,\ \ldots,\ X_n \leq x_n) = \mathrm{P}(X_1 \leq x_1)\mathrm{P}(X_2 \leq x_2)\cdots\mathrm{P}(X_n \leq x_n) \qquad \forall (x_1, x_2, \ldots, x_n) \in \overline{\mathbb{R}}^n. \tag{3.2.7}$$

Note that if we set $A_i = [X_i \leq x_i]$, then (3.2.7) is equivalent to (3.2.3), with the exception of having to check this condition for every value of $k \leq n$. Again, however, the a(3.2.7) must be true *for every value of* $(x_1, x_2, \ldots, x_n)$. Now note that if we set $x_i = \infty$, then we have $F_{X_i}(x_i) = F_{X_i}(\infty) = \mathrm{P}(X_i \leq \infty) = 1$. Therefore, we can modify (3.2.7) to be any subset of the n random variables simply by setting a certain number of values of x_i to ∞. For example, if we want to check the condition for X_1, X_2 and X_3, we set $x_i = \infty\ \forall i > 3$, thus giving us

$$F_{\boldsymbol{X}}(x_1, x_2, x_3, \infty, \infty \ldots, \infty) = F_{X_1}(x_1)F_{X_2}(x_2)F_{X_3}(x_3)F_{X_4}(\infty)F_{X_5}(\infty)\cdots F_{X_n}(\infty)$$

which is of course equivalent to

$$F_{\boldsymbol{X}}(x_1, x_2, x_3) = F_{X_1}(x_1)F_{X_2}(x_2)F_{X_3}(x_3).$$

Example 3.2.4. From Example 3.2.3 with 1 denoting heads and X_i denoting the outcome of the i^{th} toss,

$$p_{\boldsymbol{X}}(1,1,1,1,1) = p_{X_1}(1)p_{X_2}(1)p_{X_3}(1)p_{X_4}(1)p_{X_5}(1) = \frac{1}{2^5}.$$ □

Example 3.2.5. Two dice are rolled. Let the r.v. X_i denote the outcome on the i^{th} die, $i = 1,$. Consider the joint r.v. (X_1, X_2) defined over the sample space $S_1 \times S_2$:

$$p_{X_1,X_2}(x_1, x_2) = p_{X_1}(x_1) \cdot p_{X_2}(x_2) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} \quad \forall (x_1, x_2) \in S_1 \times S_2. \qquad \square$$

Example 3.2.6. Suppose the joint pdf for the r.v. (X, Y) is given by

$$f_{X,Y}(x, y) = \begin{cases} \lambda^2 e^{-\lambda(x+y)} & x, y \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

Are X and Y independent? We first compute

$$\begin{aligned} f_X(x) &= \int_0^\infty \lambda^2 e^{-\lambda(x+y)} \, dy \\ &= \lambda e^{-\lambda x} \int_0^\infty \lambda e^{-\lambda y} \, dy \\ &= \lambda e^{-\lambda x} \left[-e^{-\lambda y} \right]_{y=0}^{y=\infty} \\ &= \lambda e^{-\lambda x} (0 + 1) \\ &= \lambda e^{-\lambda x}, \quad x \geq 0, \end{aligned}$$

$$f_Y(y) = \lambda e^{-\lambda y}, \quad y \geq 0. \qquad \text{by symmetry}$$

By inspection, it can be verified that $f_{XY}(x, y) = f_X(x) f_Y(y) \ \ \forall (x, y) \in \mathbb{R}^2$. Thus, X and Y are independent. $\square$

The following theorem is stated for continuous r.v.'s, but a similar theorem can be stated and proven for discrete r.v.'s. Since the process is the same, only the continuous case will be stated and proven; we can assume the discrete case will follow.

Theorem 3.2.2. Let the r.v.'s X_1 and X_2 have the joint pdf $f_{X_1,X_2}(x_1,x_2)$. The r.v.'s X_1 and X_2 are independent iff $f_{X_1,X_2}(x_1,x_2)$ can be written as a product of a non-negative function of x_1 alone and a non-negative function of x_2 alone. That is, if

$$f_{X_1,X_2}(x_1,x_2) = g(x_1)h(x_2) \quad \forall (x_1,x_2) \in \mathbb{R}^2,$$

where $g(x_1) \geq 0$ and $h(x_2) \geq 0$. The set of points where $g(x_1) > 0$ is independent of x_2, and the set of points where $h(x_2) > 0$ is independent of x_1.

Proof:

$\Rightarrow$ If X_1 and X_2 are independent, then the theorem holds, since $f_{X_1,X_2}(x_1,x_2) = f_{X_1}(x_1)f_{X_2}(x_2) = g(x_1)h(x_2)$, where $g(x_1) \geq 0$ and $h(x_2) \geq 0$. Furthermore, the set of points where $g(x_1) > 0$ is independent of x_2, and the set of points where $h(x_2) > 0$ is independent of x_1.

$\Leftarrow$ Suppose $f_{X_1,X_2}(x_1,x_2) = g(x_1)h(x_2)$, where $g(x_1) \geq 0$ and $h(x_2) \geq 0$ and the set of points where $g(x_1) > 0$ is independent of x_2, and the set of points where $h(x_2) > 0$ is independent of x_1. Then

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} g(x_1)h(x_2)\ dx_2 = g(x_1)\int_{-\infty}^{\infty} h(x_2)\ dx_2, \tag{3.2.8}$$

$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} g(x_1)h(x_2)\ dx_1 = h(x_2)\int_{-\infty}^{\infty} g(x_1)\ dx_1, \tag{3.2.9}$$

where neither $\int_{-\infty}^{\infty} h(x_2)\ dx_2$ nor $\int_{-\infty}^{\infty} g(x_1)\ dx_1$ depend on x_1 or x_2. Furthermore, note that since the set of points where $g(\cdot) > 0$ and $h(\cdot) > 0$ does not depend on x_2 or x_1, respectively, then

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)\ dx_1dx_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1)h(x_2)\ dx_1dx_2 = \int_{-\infty}^{\infty} g(x_1)\ dx_1 \int_{-\infty}^{\infty} h(x_2)\ dx_2. \tag{3.2.10}$$

Therefore,

$$\begin{aligned} f_{X_1,X_2}(x_1,x_2) &= g(x_1)h(x_2) \\ &= g(x_1)\left[\int_{-\infty}^{\infty} g(x_1)\ dx_1 \int_{-\infty}^{\infty} h(x_2)\ dx_2\right] h(x_2) && \text{by (3.2.10)} \\ &= g(x_1)\int_{-\infty}^{\infty} h(x_2)\ dx_2 \cdot h(x_2)\int_{-\infty}^{\infty} g(x_1)\ dx_1 \\ &= f_{X_1}(x_1)f_{X_2}(x_2), && \text{by (3.2.8) and (3.2.9)} \end{aligned}$$

which shows that X_1 and X_2 are independent.

■

Example 3.2.7. Suppose

$$f_{X_1,X_2}(x_1,x_2) = \begin{cases} x_1 + x_2 & x_1, x_2 \in (0,1) \\ 0 & \text{otherwise} \end{cases}.$$

Then X_1 and X_2 are not independent, since $f(x_1,x_2)$ cannot be factored into a product of a non-negative function of x_1 and a non-negative function of x_2 alone. □

Example 3.2.8. Suppose

$$f_{X_1,X_2}(x_1,x_2) = \begin{cases} 8x_1x_2 & 0 < x_1 < x_2 < 1 \\ 0 & \text{otherwise} \end{cases}.$$

Then X_1 and X_2 are not independent, since the set of points where $g(\cdot) > 0$ depends on x_2 and the set of points where $h(\cdot) > 0$ depends on x_1. □

Example 3.2.9. Suppose

$$f_{X_1,X_2}(x_1,x_2) = \begin{cases} \lambda^2 e^{-\lambda(x_1+x_2)} & x_1, x_2 \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

Then X_1 and X_2 are independent, since we can write $f_{X_1,X_2}(x_1,x_2)$ as

$$f_{X_1,X_2}(x_1,x_2) = \begin{cases} \left(\lambda^2 e^{-\lambda x_1}\right)\left(e^{-\lambda x_2}\right) = g(x_1)h(x_2) & x_1, x_2 \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

where $g(\cdot) \geq 0$, $h(\cdot) \geq 0$, and where the domains where $g(\cdot) > 0$ and $h(\cdot) > 0$ do note depend on x_2 or x_1, respectively. □

4 Probability Laws

4.1 Discrete Distributions

4.1.1 Bernoulli

$$p_X(x) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \end{cases}.$$

p is a parameter such that $0 \leq p \leq 1$. We could write $p_X(x)$ as

$$p_X(x) = p^x(1-p)^{1-x}, \quad x \in \{0,1\}.$$

We denote this as $X \sim \text{Bernoulli}(p)$.

Example 4.1.1. Toss a coin. Let 1 denote a head and 0 denote a tail. If the coin is fair, then $p = \frac{1}{2}$. □

Example 4.1.2. Consider an urn with N balls, of which n_1 are red and $n_2 = N - n_1$ are white. Draw with replacement, and let drawing a red ball be designated as a success.

Let 1 denote a red ball drawn,
Let 0 denote a white ball drawn.

Then we have $p = \frac{n_1}{N}$ and $1 - p = \frac{n_2}{N}$. □

4.1.2 Binomial

Let X denote the number of successes in n independent Bernoulli trials:

$$X = X_1 + \cdots + X_n, \quad X_i \overset{iid}{\sim} \text{Bernoulli}(p).$$

where *iid* stands for *i*ndependently and *i*dentically *d*istributed. Our sample space is

$$S = \{(x_1 + \cdots + x_n) : \ x_i = 0, 1\}.$$

Since these are *independent* Bernoulli trials, we can multiply their respective probabilities, so that, for example,

$$\text{P}(\{(1,1,0,1,0)\}) = pp(1-p)p(1-p) = p^3(1-p)^2.$$

In general,

$$\text{P}(\{(1,1,0,1,0,\ldots,0,1)\}) = p^k(1-p)^{n-k},$$

where n is the number of trials and $k = \sum_i x_i$ is the number of successes (1's) in those trials. however, to find $\text{P}(X = k)$, we must add the above probability for *all* configurations of the zeroes and ones such that $\sum_i x_i = k$. That is,

$$\text{P}(X = k) = \text{P}(\{(x_1, \ldots, x_n) : \ \sum_{i=1}^{n} x_i = k; \ x_i = 0, 1\}) = ap^k(1-p)^{n-k},$$

where a is the number of ways to pick k successes out of n trials. However, that number is by definition $\binom{n}{k}$. Altogether, using a little change of notation,

$$p_X(x) = \binom{n}{x} p^x(1-p)^{n-x}, \quad x \in \{0, 1, \ldots, n\}.$$

Example 4.1.3. Toss a coin with probability $p = \frac{1}{2}$ for a head. Toss the coin 10 times. What is the probability that in 10 throws, one obtains 2 heads? 5 heads?

$$p_X(2) = \binom{10}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^8 = \frac{10!}{8!2!2^{10}} \approx 4.39\%.$$
$$p_X(5) = \binom{10}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^5 = \frac{10!}{5!5!2^{10}} \approx 24.61\%.$$ □

4.1.3 Hypergeometric

Draw from an urn without replacement. There are N balls, where $p \in [0,1]$ and Np is an integer:

$$100p\% \text{ are red,}$$
$$100(1-p)\% \text{ are white.}$$

Take a sample of size n without replacement. Let X denote the number of red balls drawn.

$$p_X(x) = \frac{\binom{Np}{x}\binom{N(1-p)}{n-x}}{\binom{N}{n}}, \quad x \in \{0,1,\ldots,\min(n,Np)\}.$$

Example 4.1.4. A population contains 200 articles of which 5% are defective. That is, we have $200(0.05) = 10$ defective articles. What is the probability that a sample of size 20 without replacement has exactly 3 defective articles?

$$p_X(3) = \frac{\binom{10}{3}\binom{190}{20-3}}{\binom{200}{20}} = \frac{\binom{10}{3}\binom{190}{17}}{\binom{200}{20}} \approx 5.48\%.$$

If the sample is drawn *with* replacement, then

$$p_X(3) = \binom{20}{3}(0.05)^3(0.95)^{17} \approx 5.96\%.$$ □

Note that there isn't much of a difference whether we choose with or without replacement. This is because we have a large population size $N = 200$ with respect to our sample $n = 20$. In fact, it can be shown (e.g., a variation on Ferguson (1996, pp. 201-202)) that for a given proportion p and sample size n, as $N \to \infty$, the hypergeometric distribution approaches the binomial. Thus, when N is large with respect to n, there is practically no difference between sampling with and without replacement.

Example 4.1.5. A population contains 2000 articles, of which 5% are defective. That is, there are $200(0.05) = 100$ defective articles. Draw a sample of size 4 without replacement, and let X denote the number of defective articles chosen. Since this is *without* replacement, we use the hypergeometric distribution:

$$p_X(3) = \frac{\binom{100}{3}\binom{1900}{1}}{\binom{2000}{4}} = \frac{100!4!1996!1900}{97!3!2000!} \approx 4.62 \times 10^{-4}.$$

If the sample is drawn *with* replacement, then we once again use the binomial distribution:

$$p_X(3) = \binom{4}{3}\left(\frac{100}{2000}\right)^3\left(\frac{1900}{2000}\right)^1 \approx 4.75 \times 10^{-4}. \qquad \square$$

4.1.4 Geometric

A random phenomenon obeying the geometric probability law is the number of trials required to obtain the first success (failure) in a sequence of independent, repeated Bernoulli trials in which the probability of success (failure) on each trial is p. Note that this includes the last trial (the one success (failure)). For example

$$p_X(5) = \mathrm{P}(X = 5) = \mathrm{P}(\{(0,0,0,0,1)\}) = (1-p)^4 p, \qquad p \in [0,1].$$

In general,

$$p_X(x) = p(1-p)^{x-1}, \qquad x \in \mathbb{Z}^+, \quad p \in [0,1].$$

Example 4.1.6 (Quality Control). If the probability that a toaster works is 0.9 and toasters are tested until one fails, what is the probability that the first failure o occurs on the third test?

$$p_X(3) = (0.1)(0.9)^2 = 8.10\%. \qquad \square$$

4.1.5 Poisson

A random variable X follows a *Poisson distribution* (denoted $X \sim \mathrm{P}(\lambda)$) if it has the following pmf:

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x \in \mathbb{Z}^*, \quad \lambda > 0. \tag{4.1.1}$$

The Poisson probability law has become increasingly important in recent years as more phenomena to which the law applies have been studied:

- In *physics*: The emission of electrons from filaments of a vacuum tube or from a photosensitive substance under the influence of light, radioactive decomposition.
- In *operations research*: Demands for service on cashiers, salesmen, cargo holding facilities of a port, maintenance at a machine shop, etc. Also, the role at which service is rendered.

The usual situation to which the Poisson probability law applies is in a *Poisson process*, where we are to determine the number of occurrences of some event in a time or space interval. For example, the number of planes arriving at a certain airport during a given hour, the number of organisms in a unit volume of some fluid.

By contrast, the usual situation to which the binomial probability law applies is the one in which n independent occurrences of some experiment one may determine:

- The number of trials in which a certain event occurred: E.g., the number of heads in n tosses of a coin.
- The number of trials in which a certain event did not occur.

For the Poisson process,

- We are on a timeline starting at time $t = 0$,
- One event (e.g., an earthquake) happens at a rate μ (e.g., 3 times per year),
- $N(t)$ is a discrete random variable giving the number of events that happened up to time t (in the same units as μ – e.g., 1 year).
- The probability that there are k events by time t is

$$\mathrm{P}(N(t) = k) = e^{-\mu t}\frac{(\mu t)^k}{k!} \quad k \in \mathbb{Z}^*, \quad \mu > 0. \tag{4.1.2}$$

  where μ is the average rate at which events occur per unit time or space t.

There are a couple things to note about such a process:

- It may be easy to confuse (4.1.1) with (4.1.2): Why does one use λ and the other use μt? The key to understanding this is that (4.1.1) uses a *unitless* parameter λ, whereas (4.1.2) uses a *rate* μ, which is expressed as a number of events per unit. If t is expressed in terms of the same unit, then

$$\mu t = \frac{\#}{1\ \cancel{\text{unit}}} \cdot (\#\ \cancel{\text{units}})$$

  where the units cancel each other out. Thus, μt is also unitless. We can simply think of (4.1.2) as a special case of (4.1.1), where we set $\lambda = \mu t$.
- Unfortunately, some books (notably, our suggested textbook by Ross) use λ to denote both the parameter of a Poisson distribution and the rate of a Poisson process. This is definitely confusing.
- In practice, we can simply think of (almost) every use of the Poisson distribution as a Poisson process. That is, in most applications, we will have some kind of rate μ, which will need to be multiplied by some quantity (not just time) t, as long as the units match (and thus cancel each other out). By doing this, we avoid potential confusion about what λ value to use. For example, Ross (2006, Ex. 7a, p. 162) describes typographical errors with a parameter $\lambda = \frac{1}{2}$. Within the context of the problem, this is really a Poisson process with $\mu = \frac{1 \text{ error}}{2 \text{ pages}}$ and $t = 1$ page. Not only does this alleviate any confusion as to what is happening, but it also allows us to change the problem appropriately if the quantity changes (e.g., what if we want to model typographical errors for 2 pages?).

We will say more about the Poisson process later on in this chapter.

Example 4.1.7. Observe times at which autos arrive at a toll collector's booth on a toll bridge. Suppose we know that the average rate μ of arrival of autos per minute is 1.5. The probability that x autos will arrive in a time period of length one minute is

$$p_X(x) = \frac{e^{-1.5}(1.5)^x}{x!}, \quad x \in \mathbb{Z}^*$$

whereas the probability that x autos will arrive in a half-minute period is

$$p_X(x) = \frac{e^{-1.5/2}(1.5/2)^x}{x!}. \quad x \in \mathbb{Z}^* \qquad \square$$

For the derivation of the Poisson probability law, we paraphrase from the derivation given by Feller (1968, pp. 446-448) or Ross (2006, pp. 170-172). Let X_t be the size of the population[1] at time t. Let $p_X(n;t) = \mathrm{P}(X_t = n)$. We assume the following three postulates $\forall t > 0$ and $\forall h > 0$ such that h is "small":

(1) The probability that the population will increase by 1 in a time interval from t to $t+h$ is

$$\mu h + o(h),$$

where $\mu > 0$ and $\lim_{n\to 0} \frac{o(h)}{h} = 0$. The probability of this increase does not depend on the number in the population at time t.

(2) The probability that in the time interval from t to $t+h$, the population size will change by 2 or more is

$$o(h).$$

This probability is independent of population size.

(3) The probability that the population size will.

These postulates will allow us to calculate $p_X(n;t) = \mathrm{P}(X_t = n)$.

For $n \geq 1$, the event $[X_{t+h} = n]$ can happen in any one of two mutually exclusive ways:

(a) The population at time t is n and there is no change in the population from t to $t+h$.

(b) The population at time t is $n-1$ and increases by 1 from time t to $t+h$.

Let

$$\begin{aligned}
\mathrm{P}[(\mathrm{a})] &= (\text{Probability in state } n \text{ at time } t) \cdot (\text{Probability no increase from time } t \text{ to } t+h) \\
&= p_X(n;t)\,[1 - \mu h - o(h) - o(h)] \\
\mathrm{P}[(\mathrm{b})] &= (\text{Probability in state } n-1 \text{ at time } t) \cdot (\text{Probability increase of 1 from time } t \text{ to } t+h) \\
&= p_X(n-1;t)\,[\mu h + o(h)]
\end{aligned}$$

Therefore, for $n \geq 1$,

$$\mathrm{P}(X_{t+h} = n) = p_X(n;t+h) = p_X(n;t)[1 - \mu h - 2o(h)] + p_X(n-1;t)[\mu h + o(h)] \tag{4.1.3}$$

while for $n = 0$,

$$\mathrm{P}(X_{t+h} = 0) = p_X(o;t+h) = p_X(o;t)[1 - \mu h - 2o(h)]. \tag{4.1.4}$$

Equation (4.1.3) may be written as

$$\frac{p_X(n;t+h) - p_X(n;t)}{h} = -\mu p_X(n;t) + \mu p_X(n-1;t) + \frac{o(h)[-2p_X(n;t) + p_X(n-1;t)]}{h}$$

Taking the limit as $h \to 0$, we have

$$\lim_{h\to 0} \frac{p_X(n;t+h) - p_X(n;t)}{h} = \frac{\partial}{\partial t} p_X(n;t) = -\mu p_X(n;t) + \mu p_X(n-1;t) \tag{4.1.5}$$

[1] The population could be particles emitted by a radioactive source, biological organisms of a given kind present in a certain environment, persons waiting in a queue for service, etc.

since $\lim_{h\to 0} \frac{o(h)}{h} = 0$. If we take the limit of (4.1.4) as $h \to 0$, we have

$$\frac{\partial}{\partial t} p_X(0;t) = -\mu p_X(0;t). \tag{4.1.6}$$

We now must solve (4.1.5) and (4.1.6) for $p_X(n;t)$, which requires differential equations. The questions of existence and uniqueness of solutions of these partial differential equations will not be discussed here.

For $n = 0$ under the assumption that $p_X(0;0) = 1$, (4.1.6) has the solution

$$p_X(0;t) = e^{-\mu t}.$$

For $n = 1$ under the assumption that $p_X(1;0) = 0$, (4.1.5) has the solution[2]

$$p_X(1;t) = \mu e^{-\mu t} \int_0^t e^{\mu x} p_X(0;x)\, dx = \mu e^{-\mu t} t = \mu t e^{-\mu t}.$$

By induction under the assumption that $p_X(n;0) = 0$ for $n > 0$, we obtain

$$p_X(n;t) = \frac{e^{-\mu t}(\mu t)^n}{n!}, \qquad x \in \mathbb{Z}^*.$$

Example 4.1.8. Suppose a retailer discovers that the number of items of a certain kind demanded by customers in a given period obeys a Poisson probability law with known parameter λ. What stock k of this item should the retailer have on hand at the beginning of the time period in order to have a probability of at least 0.99 that he will be able to supply immediately all customers who demand the item under consideration?

Let X denote the number of items demanded in a given period. Then

$$\mathrm{P}(X \le k) = \sum_{x=0}^{k} \frac{e^{-\lambda}\lambda^x}{x!} \ge 0.99 \quad \Leftrightarrow \quad \mathrm{P}(X > k) = \sum_{x=k+1}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \le 0.01.$$

To find k, use a computer or see the Poisson tables for the particular value of λ. For instance,

$$\begin{aligned}
\lambda = 1 \quad &\Rightarrow \quad \sum_{x=5}^{\infty} \frac{e^{-1}1^x}{x!} = 0.0037, \ \sum_{x=4}^{\infty} \frac{e^{-1}1^x}{x!} = 0.0190 && \Rightarrow k = 4.\\
\lambda = 3 \quad &\Rightarrow \quad \sum_{x=9}^{\infty} \frac{e^{-3}3^x}{x!} = 0.0038, \ \sum_{x=8}^{\infty} \frac{e^{-3}3^x}{x!} = 0.0119 && \Rightarrow k = 8.\\
\lambda = 20 \quad &\Rightarrow \quad \sum_{x=32}^{\infty} \frac{e^{-20}20^x}{x!} = 0.0081 && \Rightarrow k = 31.
\end{aligned}$$

□

Theorem 4.1.1. Let X have a binomial distribution with parameters n and p. If $\lambda = np$, then $p = \frac{\lambda}{n}$. Fix λ and let n and p vary.

$$\lim_{n\to\infty} \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} = \frac{e^{-\lambda}\lambda^x}{x!}.$$

[2]Recall that an integrating factor is used to solve a differential equation $\frac{dy}{dt} + Py = Q$. The integrating factor is $\rho = \exp(\int P\, dt)$, and the solution is $\rho y = \int \rho Q\, dt + C$.

Proof: We know that

$$\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} = \frac{1}{x!}\lambda^x\left(1-\frac{\lambda}{n}\right)^{n-x}\frac{n(n-1)\cdots(n-x+1)}{n^x}.$$

We know that

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda} \quad \text{by definition}$$
$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x} = 1$$
$$\lim_{n\to\infty}\frac{\lambda^x}{x!} = \frac{\lambda^x}{x!}$$
$$\lim_{n\to\infty}\frac{n}{n}\frac{(n-1)}{n}\frac{(n-2)}{n}\cdots\frac{(n-x+1)}{n} = 1$$

Therefore,

$$\lim_{n\to\infty}\left[\frac{1}{x!}\lambda^x\left(1-\frac{\lambda}{n}\right)^{n-x}\frac{n(n-1)\cdots(n-x+1)}{n^x}\right] = \frac{\lambda^x}{x!}e^{-\lambda}\cdot 1\cdot 1.$$

■

In the above theorem, we could also write the approximation as

$$\lim_{n\to\infty}\binom{n}{k}p^k(1-p)^{n-k} = \frac{e^{-np}(np)^k}{k!}.$$

From this result, we can deduce that we can use the *Poisson approximation to the binomial* if $p \leq 0.1$, n is large and np is not too large. For example if $n = 100$ and $p = 0.01$,

x	Binomial(100,0.01)	Poisson, $np = 1$
0	0.366	0.368
1	0.370	0.368
2	0.185	0.184
3	0.061	0.061

Example 4.1.9. Suppose it is well known that the probability that an item is produced by a certain machine will be defective is 0.1. Find the probability that a sample of 10 items, selected at random from the output of the machine, will contain no more than one defective item.

$$\text{Binomial: } \mathrm{P}(X \leq 1) = \binom{10}{0}(0.1)^0(0.9)^{10} + \binom{10}{1}(0.1)^1(0.9)^9 = 73.61\%.$$

$$\text{Poisson: } \mathrm{P}(X \leq 1) = \frac{e^{-1}1^0}{0!} + \frac{e^{-1}1^1}{1!} = 73.58\%.$$

where for the Poisson, we set $\lambda = np = 1$. □

Example 4.1.10. Assume the number of live polio cells in a certain Polio vaccine follows a Poisson probability law with $\mu = 1$ cell per milliliter. let X denote the number of cells per milliliter.

$$\mathrm{P}(X = x) = \frac{e^{-1}1^x}{x!}, \quad x \in \mathbb{Z}^*.$$

Let Y denote the number of cells per 10 milliliters: $\mu t = 1 \cdot 10 = 10$.

$$\mathrm{P}(Y = y) = \frac{e^{-10}10^y}{y!}, \quad y \in \mathbb{Z}^*.$$ □

4.2 Continuous Distributions

4.2.1 Uniform Distribution

Assume that the density function is non-zero only for values of $x \in (a, b)$ and that the pdf is defined so that if B is any interval contained in (a, b), then

$$\mathrm{P}(B) = \frac{\text{length of } B}{\text{length of } (a,b)} = \frac{\text{length of } B}{b-a}.$$

The uniform distribution is an extension of the notion of a finite sample space S of size n, where

$$\mathrm{P}(\{x_i\}) = \frac{1}{n} \quad \forall x_i \in S.$$

For the uniform distribution, the distribution function $F_X(x) = \mathrm{P}(X \leq x)$ is given by

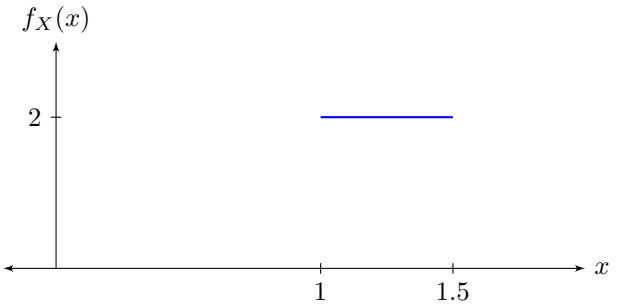
$$F_X(x) = \begin{cases} 0 & x \leq a \\ \dfrac{x-a}{b-a} & a < x < b \\ 1 & x \geq b \end{cases}.$$

the pdf is given by

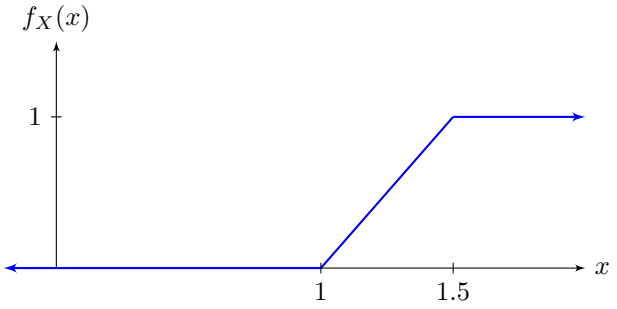
$$f_X(x) = F_X'(x) = \frac{1}{b-a}, \quad a < x < b.$$

Example 4.2.1. If X is a uniform rv with $(a, b) = (1, 1.5)$, then

$$f_X(x) = \frac{1}{1.5 - 1} = 2, \qquad 1 < x < 1.5,$$



$$F_X(x) = \begin{cases} 0 & x < 1 \\ 2(x-1) & 1 \le x \le 1.5 \\ 1 & x > 1.5 \end{cases}.$$



Furthermore,

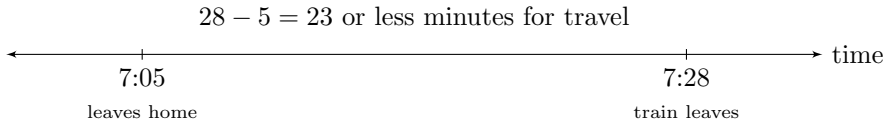
$$\mathrm{P}(0 < X < 1.2) = \int_1^{1.2} 2 \, dx = [2x]_1^{1.2} = 2.4 - 2 = 0.4,$$

$$\begin{aligned} \mathrm{P}(-3 < X < 6) &= F_X(6) - F_X(3) = 1 - 0 = 1, \\ \text{or } &= \int_{-3}^{6} f_X(x) \, dx = \int_1^{1.5} 2 \, dx = [2x]_1^{1.5} = 3 - 2 = 1. \end{aligned}$$

□

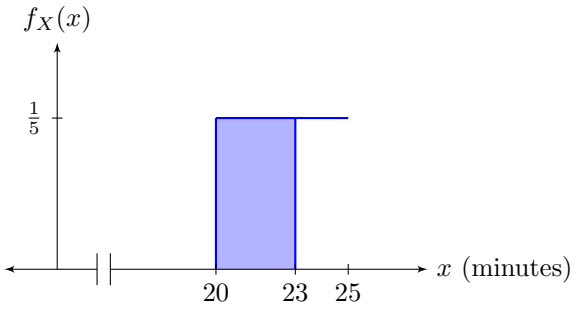
Example 4.2.2. The time, measured in minutes, required by a certain man to travel from his home to the train station is a random phenomenon obeying a uniform probability law over the interval from 20 to 25 minutes. If he leaves home promptly at 7:05 am, what is the probability that he will catch a train that leaves the station promptly at 7:28 am?

He will catch the train if he arrives at the station on or before 7:28 am. Let the rv X denote the time in minutes it takes to travel from home to the station. By hypothesis, X has a uniform $(20, 25)$ distribution.



He thus catches the train if $X \leq 23$ minutes.

$$\mathrm{P}(X \leq 23) = \int_{20}^{23} \frac{1}{5}\, dx = \frac{3}{5} = 60\%.$$



Note that $\mathrm{P}(X \leq 23) = F_X(23)$. □

4.2.2 Exponential Distribution

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \quad \lambda > 0.$$

This is denoted as $X \sim \mathrm{Exp}(\lambda)$. Phenomena which obey this law are *life lengths* of almost anything: light bulbs, machines, and even humans (from ages 2 to 18, if they don't drive). λ is the *failure rate*, expressed as the number of failures per unit time.

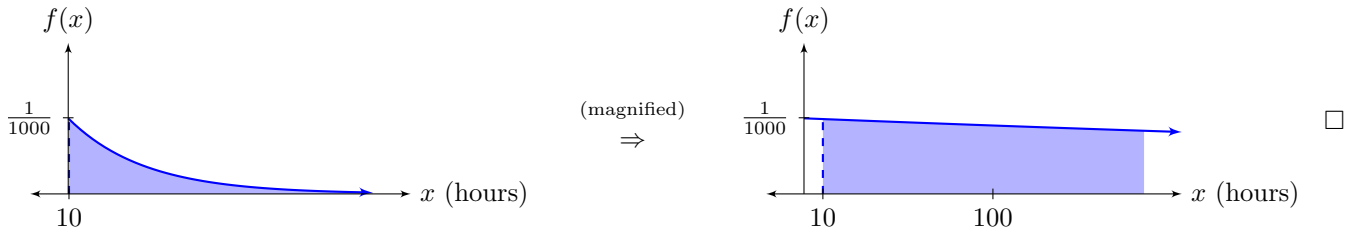
Example 4.2.3. Consider a radar set of the type whose failure law is exponential. If radar sets of this type have a failure rate of $\lambda = 1$ set per 1000 hours, find the length of time t such that the probability is 0.99 that a set will operate satisfactorily for a time greater than t.

Let T denote the time to failure. Then

$$\mathrm{P}(T > t) = 0.99 = \int_t^\infty \frac{1}{1000} e^{-x/1000}\, dx = \left[-e^{-x/1000}\right]_t^\infty = e^{-t/1000} = 0.99.$$

Taking the log of both sides gives

$$\frac{-t}{1000} = \ln(0.99) \quad \Leftrightarrow \quad t = -1000 \ln(0.99) = 10 \text{ hours}.$$



□

Theorem 4.2.1. Given a Poisson process with parameter μ, designate as time 0 the time at which we start observing the process. Let T be the time which passes until the first event occurs. T has an exponential distribution with parameter μ.

Proof: Our diagram is the following:

If $t < 0$, then $\mathrm{P}(T \leq t) = 0$. Now let $t \geq 0$ be given. then

$$\mathrm{P}(T > t) = \text{probability that no events occur in } (0, t] = \frac{e^{-\mu t}(\mu t)^0}{0!} = e^{-\mu t}.$$

Therefore,

$$F_T(t) = 1 - \mathrm{P}(T > t) = 1 - e^{-\mu t} \quad \Rightarrow \quad f_T(t) = F_T'(t) = \mu e^{-\mu t} \text{ for } t > 0.$$

It follows that the interarrival times between two Poisson events has an exponential distribution. ■

It also can be proven that if the interarrival times between *any* two arrivals (must hold for all arrival times) of a process have an exponential distribution with parameter μ, then the process is a Poisson process with an average number of arrivals per unit time $= \mu$. That is, the above theorem is an if and only if statement.

Notationally, let $Y_1, Y_2, \ldots$ denote the inter-arrival times of the events. Specifically, let Y_k denote the time between the $k-1^{th}$ and k^{th} events. First of all, note that

$$[Y_1 > t] = [N(t) = 0] = [N(t) < 1] \quad \Rightarrow \quad \mathrm{P}(Y_1 > t) = \mathrm{P}(N(t) < 1).$$

This is proven by thinking about what these two events mean, in plain English:

- $[Y_1 > t]$: "The time between our starting time ($t = 0$) and the first event is greater than t,"
- $[N(t) = 0]$: "No events happened between the starting time and time t."

Likewise, using the same logic, we have that

$$[Y_1 < t] = [N(t) > 0] = [N(t) \geq 1] \quad \Rightarrow \quad \mathrm{P}(Y_1 < t) = \mathrm{P}(N(t) \geq 1). \tag{4.2.7}$$

The event $[Y_1 = t]$ is of no consequence, since $\mathrm{P}(Y_1 = t) = 0$ (since Y_1 is a continuous random variable).

Secondly, define $W_k = Y_1 + Y_2 + \cdots + Y_k$. Then note that

$$[W_k > t] = [N(t) < k] \quad \Rightarrow \quad \mathrm{P}(W_k > t) = \mathrm{P}(N(t) < k).$$

since once again in plain English

- $[W_k > t]$: "The time begin our starting time and the k^{th} event is greater than t,"
- $[N(t) < k]$: "Fewer than k events happened between the starting time and time t."

Likewise,

$$[W_k < t] = [N(t) \geq k] \quad \Rightarrow \quad \mathrm{P}(W_k < t) = \mathrm{P}(N(t) \geq k).$$

As with Y_1, the event $[W_k = t]$ is of no consequence.

Now, what can we say about the event $[N(t) = k]$?

$$\begin{aligned}[N(t) = k] &= [\text{there are } k \text{ events between time 0 and time } t] \\ &= [\text{the } k^{th} \text{ event happened before time } t] \cap [\text{the } k+1^{th} \text{ event happened after time } t] \\ &= [W_k < t] \cap [W_{k+1} > t] \\ &= [W_k < t] \cap [Y_{k+1} > t - W_k].\end{aligned}$$

so that

$$\mathrm{P}(N(t) = k) = \mathrm{P}([W_k < t] \cap [W_{k+1} > t]).$$

Later on, we'll learn how to evaluate the probability on the right of this equation. For now, it's enough to just understand the logic behind it.

How is Y_1 distributed? By (4.1.2), we have that

$$F_{Y_1}(t) = \mathrm{P}(Y_1 \leq t) = \mathrm{P}(N(t) > 0) = 1 - e^{-\mu t} \quad \Rightarrow \quad f_{Y_1}(t) = \frac{dF_{Y_1}(t)}{dt} = \mu e^{-\mu t}, \ t > 0.$$

Thus, Y_1 is distributed as an exponential distribution with parameter μ. Furthermore, because of the *memoryless property* of a Poisson process (part of the first assumption[3]), the same can be said for Y_k for all values of k. All in all,

The inter-arrival times of a Poisson process are exponentially distributed with parameter μ.

Example 4.2.4. In the morning, students enter the STAT/MATH 394 class at a rate of 1 for every 3 minutes.

1. What is the probability that no one enters between 8:15 and 8:20?

   Solution:
   Here we have $\mu = \frac{1}{3 \text{ minutes}}$ and $t = 5$ minutes, so that, using (4.1.2), $\mu t = \frac{5}{3}$ and so
   $$\mathrm{P}(N(5) = 0) = e^{-5/3} \approx 18.89\%.$$ □

2. What is the probability that at least 4 students enter the classroom during that time?

   Solution:
   Again with $\mu t = \frac{5}{3}$, we have by (4.1.2) again that

   $$\mathrm{P}(N(5) \geq 4) = 1 - \sum_{k=0}^{3} \mathrm{P}(N(5) = k) = 1 - \sum_{k=0}^{3} e^{-5/3} \frac{(5/3)^k}{k!} \approx 8.83\%.$$ □

[3] "The probability of this increase does not depend on the number in the population at time t,"

Example 4.2.5. Suppose that the number of accidents occurring on a highway each day is a Poisson random variable with parameter $\mu = 3$.

1. Find the probability that 3 or more accidents occur today.

   Solution:
   First of all, conceptually, note that while we are dealing with the regular Poisson distribution here (4.1.1), we are implicitly modeling a Poisson process (4.1.2). That is, we can think of $\mu = \frac{3}{1 \text{ day}}$ and $t = 1$ day, so that

   $$\lambda = \mu t = \frac{3}{1 \cancel{\text{day}}} \cdot 1 \cancel{\text{day}} = 3.$$

   Therefore, using (4.1.2),

   $$\mathrm{P}(N(1) \geq 3) = 1 - \sum_{k=0}^{2} \mathrm{P}(N(1) = k) = 1 - \sum_{k=0}^{2} e^{-3} \frac{3^k}{k!} \approx 57.68\%. \qquad \square$$

2. Find the probability that 3 or more accidents occur today, given that at least 1 accident occurs today.

   Solution:
   Here we simply use the formula for a conditional distribution:

   $$\begin{aligned} \mathrm{P}(N(1) \geq 3 | N(1) \geq 1) &= \frac{\mathrm{P}\left([N(1) \geq 3] \cap [N(1) \geq 1]\right)}{\mathrm{P}(N(1) \geq 1)} && \text{definition of conditional prob} \\ &= \frac{\mathrm{P}\left(N(1) \geq 3\right)}{\mathrm{P}(N(1) \geq 1)} && \text{since } [N(1) \geq 3] \subseteq [N(1) \geq 1] \\ &= \frac{1 - \sum_{k=0}^{2} e^{-3} \frac{3^k}{k!}}{1 - e^{-3}} \\ &\approx 60.70\%. \end{aligned}$$ $\square$

Example 4.2.6. Customer arrive at McDonald's according to a Poisson process with parameter $\mu = 0.5$/minute. Let T denote the time of arrival of the next customer. Starting at time 0, what is the probability that the next customer arrives more than 2 minutes from now?

$$\mathrm{P}(T > 2 \text{ minutes}) = \int_2^\infty 0.5 e^{-0.5t} \, dt = e^{-1} \approx 36.79\%.$$

What is the probability that the next customer arrives in less than 1 minute?

$$\mathrm{P}(T < 1 \text{ minute}) = \int_0^1 0.5 e^{-0.5t} \, dt = 1 - e^{-1/2} \approx 39.35\%. \qquad \square$$

Theorem 4.2.2 (Memorylessness). If $T \sim \mathrm{Exp}(\lambda)$ and $a, b > 0$, then $\mathrm{P}(T > a + b \mid T > a) = \mathrm{P}(T > b)$.

Proof:

$$\mathrm{P}(T > a+b \mid T > a) \stackrel{\text{def}}{=} \frac{\mathrm{P}(T > a+b)}{\mathrm{P}(T > a)} = \frac{\displaystyle\int_{a+b}^{\infty} \lambda e^{-\lambda t}\, dt}{\displaystyle\int_{a}^{\infty} \lambda e^{-\lambda t}\, dt} = \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} = e^{-\lambda b} = \mathrm{P}(T > b).$$

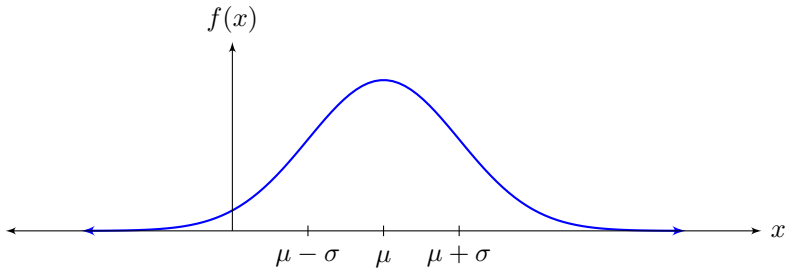
■

4.2.3 Normal Distribution

This is also known as the *Gaussian*, *LaPlace* or *bell-shaped* distribution.

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \mathbb{R}; \quad \mu \in \mathbb{R}; \quad \sigma > 0.$$

This is denoted as $X \sim \mathrm{N}(\mu, \sigma^2)$ or $X \sim \mathrm{N}(\mu, \sigma)$. To avoid confusion, we will state whether σ is squared or not, as in $X \sim \mathrm{N}(5, \sigma^2 = 3)$ or $X \sim \mathrm{N}(-2, \sigma = 5)$.



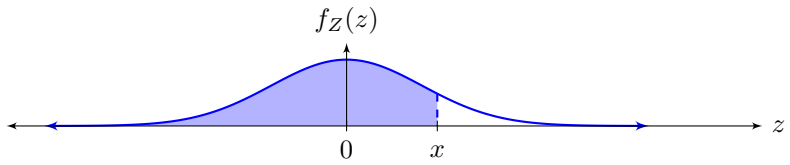
The normal distribution has been important in probability theory since the 18^{th} century. The distribution was first encountered in the work of Abraham DeMoivre (1667-1754) in 1733 as a means of approximating the distribution function of the binomial probability law for large n. This distribution is important for the fact that under various conditions (studied in detail later), it closely approximates many other probability distributions – not just the binomial.

There are physical phenomena which obey the normal probability law exactly. An example is the speed S of a molecule of mass M in a gas at absolute temperature T. According to Maxwell's laws of velocity,

$$S \sim \mathrm{N}(0, \sigma^2 = M/kT), \quad k = \textit{Boltzman's constant}.$$

The normal distribution function is tabulated for the case when $\mu = 0$ and $\sigma^2 = 1$. Since this is used often, the distribution function has its own notation, $\Phi(\cdot)$:

$$\Phi(t) = F_Z(t) = \mathrm{P}(Z \le t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$$



Area $\Phi(x)$ under the standard normal curve to the left of x

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Example 4.2.7. Using the table above, we have

$$\Phi(0.39) = 0.6517, \quad \Phi(2.99) = 0.9986. \qquad \square$$

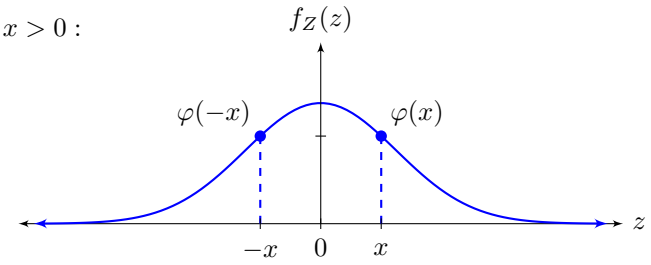
The density function is denoted by

$$\varphi(z) = f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad z \in \mathbb{R}.$$

There are two properties owing to to symmetry[4]:

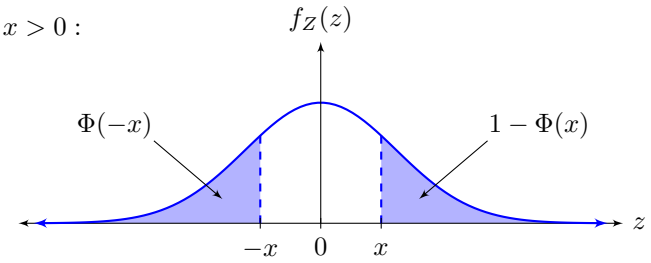
$$\varphi(x) = \varphi(-x) \quad \forall x \in \mathbb{R}$$

For $x > 0$:



$$\Phi(-x) = 1 - \Phi(x) \quad \forall x \in \mathbb{R}$$

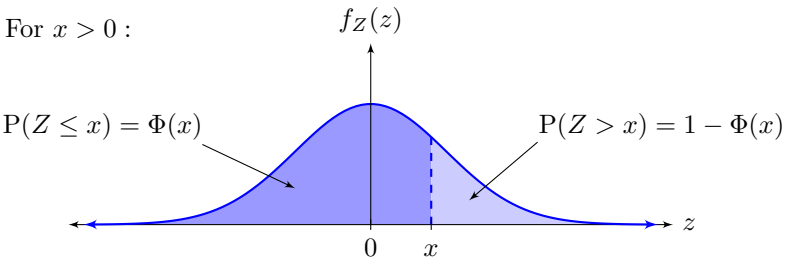
For $x > 0$:



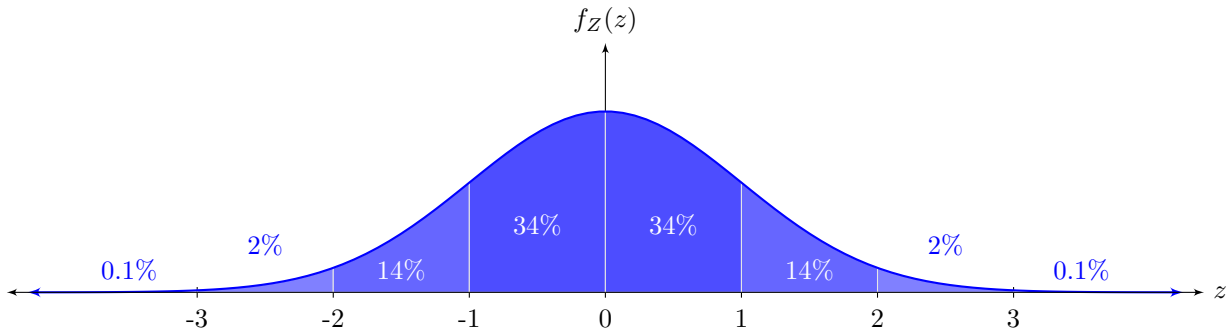
Note that, as with any distribution function $F_X(\cdot)$,

$$\mathrm{P}(Z > x) = 1 - \mathrm{P}(Z \leq x) = 1 - \Phi(x)$$

For $x > 0$:



Furthermore, as a rough guide to the table, we have this:



[4]The graphs illustrate these for $x > 0$, but the formulas are also true for $x \leq 0$.

Theorem 4.2.3. Let $X \sim N(\mu, \sigma^2)$. Then

$$\mathrm{P}(a < X < b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Proof:

$$\begin{aligned}
\mathrm{P}(a < X < b) &= \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}\, dx \\
&= \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \frac{1}{\sigma\sqrt{2\pi}} e^{-(\mu+\sigma y-\mu)^2/2\sigma^2}\, \sigma dy && y = \tfrac{x-\mu}{\sigma} \;\Leftrightarrow\; x = \mu + \sigma y \;\Leftrightarrow\; dx = \sigma dy \\
&= \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy \\
&= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). && \text{definition of } \Phi(\cdot)
\end{aligned}$$

■

Example 4.2.8. Suppose $X \sim \mathrm{N}(-2, \sigma = 3)$. What is $\mathrm{P}(-3 < X < 6)$?

$$\begin{aligned}
\mathrm{P}(-3 < X < 6) &= \Phi\left(\frac{6-(-2)}{3}\right) - \Phi\left(\frac{-3-(-2)}{3}\right) \\
&= \Phi\left(\frac{8}{3}\right) - \Phi\left(-\frac{1}{3}\right) \\
&\approx \Phi(2.667) - \Phi(-0.333) \\
&= \Phi(2.667) - (1 - \Phi(0.333)) && \text{since } \Phi(-x) = 1 - \Phi(x) \\
&\approx 0.9962 - 1 + 0.6293 && \text{from the table} \\
&= 62.55\%.
\end{aligned}$$

□

Example 4.2.9. An astronomer is interested in measuring, in light years, the distance from his observatory to a distant star. Although the astronomer has a measuring technique, he know that, due to changing atmospheric conditions and normal error, each time a measurement is made it will not yield the exact distance, but merely an estimate of μ. Assume a measurement X is normally distributed with mean d, the true distance, and a variance = 4 light years. We want to determine $\mathrm{P}(|X - d| < 1 \text{ light year})$.

$$
\begin{aligned}
\mathrm{P}(|X-d|<1 \text{ light year}) &= \mathrm{P}(-1<X-d<1) \\
&= \mathrm{P}(d-1<X<d+1) \\
&= \Phi\left(\frac{d+1-d}{\sqrt{4}}\right) - \Phi\left(\frac{d-1-d}{\sqrt{4}}\right) && \text{Note that } \sigma=\sqrt{4}=2 \\
&= \Phi\left(\frac{1}{2}\right) - \Phi\left(-\frac{1}{2}\right) \\
&= \Phi\left(\frac{1}{2}\right) - \left(1-\Phi\left(\frac{1}{2}\right)\right) && \text{since } \Phi(-x)=1-\Phi(x) \\
&= 2\Phi\left(\frac{1}{2}\right) - 1 \\
&\approx 2(0.6915) - 1 && \text{from the table} \\
&= 38.30\%. && \square
\end{aligned}
$$

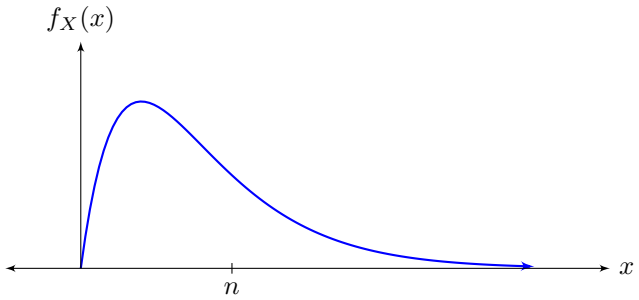
4.2.4 Chi-square Distribution

$$f_X(x) = \frac{x^{n/2-1}}{2^{n/2}\Gamma(n/2)} e^{-x/2}, \quad x>0; \quad n \in \mathbb{Z}^+.$$

This is denoted as $X \sim \chi^2_n$ and is known as the *Chi-square distribution with n degrees of freedom*. The pdf above uses the Gamma function $\Gamma(\cdot)$, defined as

$$\Gamma(n) = \int_0^\infty x^{n-1} e^{-x}\, dx, \quad n>0.$$

It can be shown (not here) that $\Gamma(n+1) = n\Gamma(n)$ and $\Gamma(1/2) = \sqrt{\pi}$. Furthermore, if $n \in \mathbb{Z}^+$, then $\Gamma(n) = (n-1)!$.

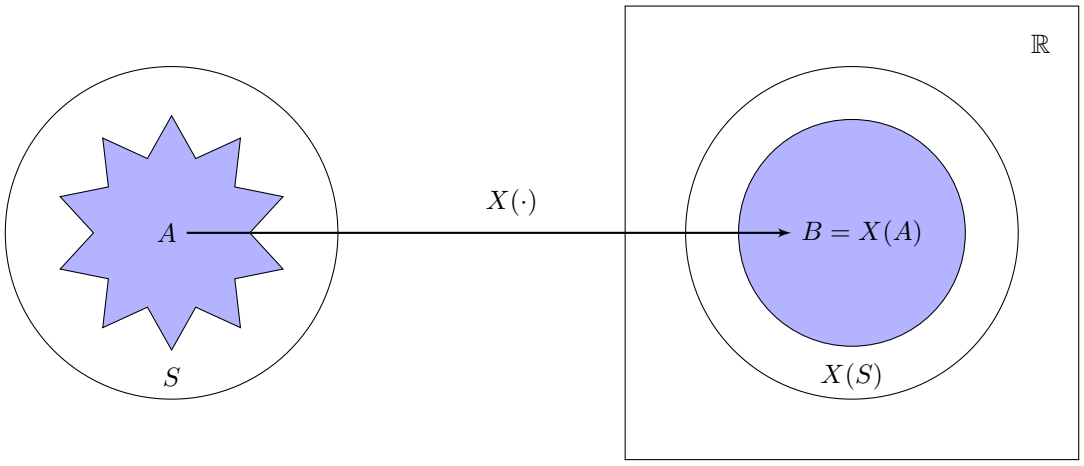


Tables are available for the distribution function $F_X(\cdot)$.

5 Functions of Random Variables

5.1 Functions of One Random Variable

Let X be a random variable. Then X is a function from the sample space S to the real numbers $\mathbb{R}$, as described in Definition 2.7.1:

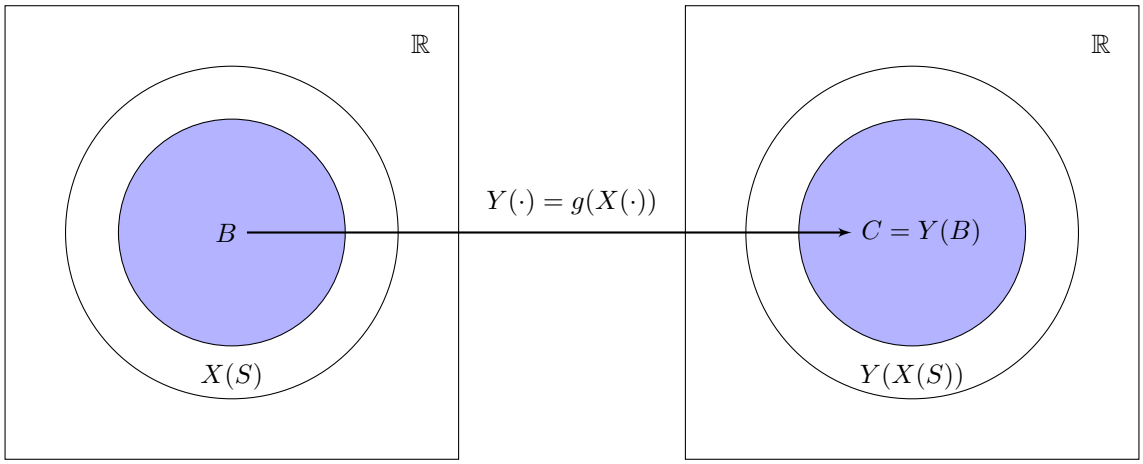


$\mathrm{P}(\cdot),\ \Psi_S$ $\qquad$ $\mathrm{P}_X(\cdot),\ \Psi_X$

We define

$$\mathrm{P}_X(B) = \mathrm{P}(A) \ \ \forall B \in \Psi_X,$$

where $A = X^{-1}(B)$ is the inverse image of B in S. We now define a *new* random variable Y, where $Y = g(X)$ and $g(\cdot)$ is a real-valued function:



$\mathrm{P}_X(\cdot),\ \Psi_X$ $\qquad$ $\mathrm{P}_Y(\cdot),\ \Psi_Y$

$\forall C \in \Psi_Y$, $\exists$ a *pre-image* $B \in \Psi_X$ which is mapped into C by Y. We then define

$$\mathrm{P}_Y(C) = \mathrm{P}_X(B) = \mathrm{P}_X(X \in g^{-1}(C)) = \mathrm{P}_X(X:\ g(x) \in C) \ \ \forall C \in \Psi_Y.$$

We will first consider the 1-1 transformation $g(X) = aX + b$, where $a > 0$ and $b \in \mathbb{R}$. We know the distribution of X and we want to determine the distribution of $Y = g(X)$.

Example 5.1.1. Consider a roulette game. There are 38 divisions:

$$18 \text{ red}, \quad 18 \text{ black}, \quad 2 \text{ grey}.$$

Assume that the divisions are equally likely to come up. We bet on red, so that if red comes up, the player doubles her stake. We thus define the random variable

$$X = \begin{cases} 0 & \text{if a non-red occurs} \\ 1 & \text{if a red occurs} \end{cases}.$$

Thus,

$$\begin{aligned} \mathrm{P}_X(\{0\}) &= \mathrm{P}(X = 0) = \frac{20}{38}, \\ \mathrm{P}_X(\{1\}) &= \mathrm{P}(X = 1) = \frac{18}{38}. \end{aligned}$$

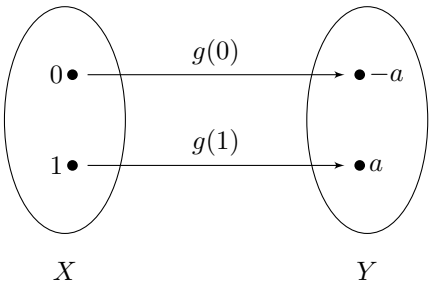
Thus, X has a Bernoulli distribution with parameter $p = \frac{18}{38}$, so that

$$p_X(x) = \left(\frac{18}{38}\right)^x \left(\frac{20}{38}\right)^{1-x}, \quad x = 0, 1.$$

Now let Y be the financial outcome of a player who bets $\$a$. What is the distribution of Y? First let's form an equation $Y = g(X)$ that gives Y as a function of X:

$$Y = 2aX - a.$$

Given this equation, the next step is to find all possible values of Y, which give the *range* of Y. Inspection shows that possible values of Y are a and $-a$, which makes Y a Bernoulli rv:



$$\Rightarrow \qquad \begin{aligned} p_Y(-a) &= p_X(0) = \frac{20}{38}, \\ p_Y(a) &= p_X(1) = \frac{18}{38}. \end{aligned}$$

Since there is a 1-1 relationship here, it is easy to determine $p_Y(\cdot)$. However, we can also figure out those values mathematically:

$$p_Y(y) = \mathrm{P}_Y(Y = y) = \mathrm{P}_X(2aX - a = y) = \mathrm{P}_X\left(X = \frac{y + a}{2a}\right) = p_X\left(\frac{y + a}{2a}\right), \quad y = -a, a.$$

Therefore,

$$\begin{aligned} p_Y(-a) &= p_X\left(\frac{-a + a}{2a}\right) = p_X(0) = \frac{20}{38}, \\ p_Y(a) &= p_X\left(\frac{a + a}{2a}\right) = p_X(1) = \frac{18}{38}, \end{aligned}$$

which agree with our previous results. □

5.1.1 Discrete Case: Linear Transformation

Suppose we have

$$Y = g(X) = aX + b.$$

Then just as we showed in the above example,

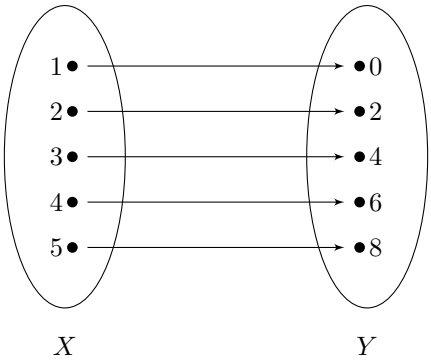
$$p_Y(y) = \mathrm{P}_Y(Y = y) = \mathrm{P}_X(aX + b = y) = \mathrm{P}_X\left(X = \frac{y-b}{a}\right) = p_X\left(\frac{y-b}{a}\right), \quad \text{where } \frac{y-b}{a} \in \mathbb{R}.$$

Example 5.1.2. Let X have the pmf

$$p_X(x) = x/15, \quad x \in \{1, 2, 3, 4, 5\}.$$

Determine the pmf for $Y = 2X - 2$.

First note that possible values of Y are $\{0, 2, 4, 6, 8\}$.



Thus, by the above,

$$p_Y(y) = p_X\left(\frac{y+2}{2}\right) = \frac{y+2}{30}, \qquad \frac{y+2}{2} \in \{1, 2, 3, 4, 5\} \quad \Leftrightarrow \quad y \in \{0, 2, 4, 6, 8\}. \qquad \square$$

5.1.2 Continuous Case: Linear Transformation

Once again, suppose we have

$$Y = g(X) = aX + b.$$

Then

$$F_Y(y) = \mathrm{P}_Y(Y \le y) = \mathrm{P}_X(aX + b \le y) = \mathrm{P}_X\left(X \le \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right)$$

and thus

$$f_Y(y) = F_Y'(y) = F_X'\left(\frac{y-b}{a}\right) \cdot \frac{\partial}{\partial y}\left(\frac{y-b}{a}\right) = f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{a}.$$

by the chain rule.

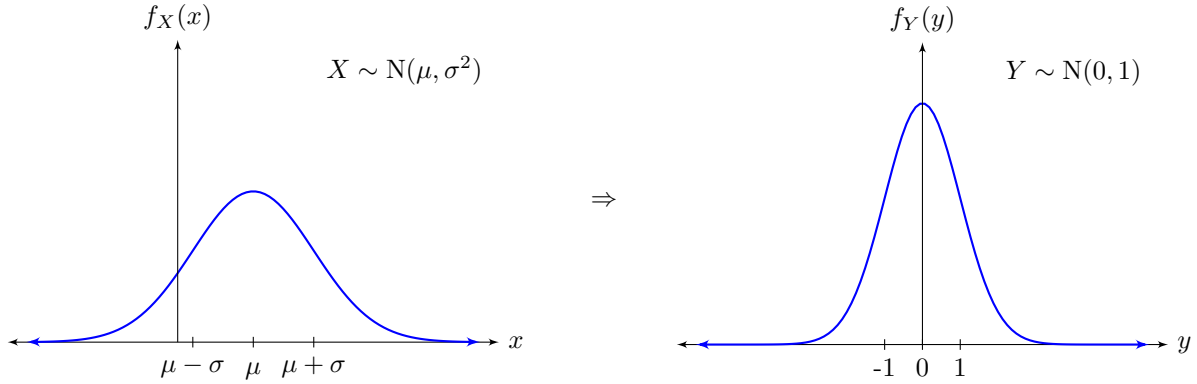
Example 5.1.3. Assume $X \sim \mathrm{N}(\mu, \sigma^2)$. Then

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \mathbb{R}; \quad \mu \in \mathbb{R}; \quad \sigma > 0.$$

Suppose $Y = \frac{X-\mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma} = aX + b$. What is the distribution of Y? By the above,

$$\begin{aligned} f_Y(y) &= \sigma f_X\left(\frac{y + \mu/\sigma}{1/\sigma}\right) \\ &= \sigma f_X(\sigma y + \mu) \\ &= \sigma \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-(\sigma y + \mu - \mu)^2/2\sigma^2} \\ &= \frac{1}{\sqrt{2\pi}} e^{-y^2/2}, \quad y \in \mathbb{R}. \end{aligned}$$

Thus, $Y \sim \mathrm{N}(0, 1)$.



Note that the distribution is not complete without also stating a range for y. □

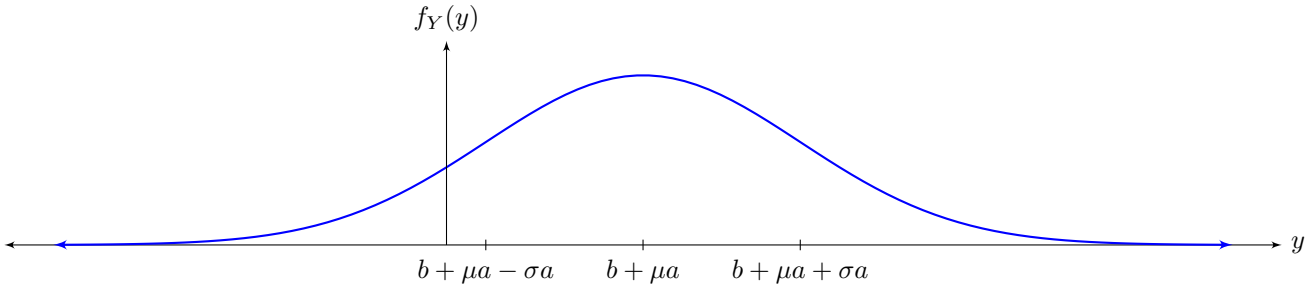
Example 5.1.4. Let X denote the molar concentration of a given compound, and let Y denote the absorbance at a given wavelength:

$$Y = aX + b,$$

where a is a constant which depends on wavelength of a compound. The compound is dissolved in a solvent. The constant b depends on the solvent. Suppose $X \sim \mathrm{N}(\mu, \sigma^2)$. What is the distribution of Y?

$$f_Y(y) = \frac{1}{a} f_Y\left(\frac{y-b}{a}\right) = \frac{1}{a\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\left(\frac{y-b}{a} - \mu\right)^2\right) = \frac{1}{a\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2a^2\sigma^2}(y - b - a\mu)^2\right).$$

Thus, Y is normal with mean $b + \mu a$ and variance $\sigma^2 a^2$.



□

Example 5.1.5. If $X \sim \mathrm{N}(\mu, \sigma^2)$, then $Y = \frac{X-\mu}{a} \sim \mathrm{N}(0,1)$ by Example 5.1.3. Since $X = \sigma Y + \mu$,

$$F_X(x) = \mathrm{P}_X(X \le x) = \mathrm{P}_Y(\sigma Y + \mu \le x) = \mathrm{P}_Y\left(Y \le \frac{x-\mu}{\sigma}\right) = F_Y\left(\frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R}.$$

For example, if $X \sim \mathrm{N}(2, \sigma^2 = 4)$, then $\mu = 2$ and $\sigma = 2$, and $F_X(x) = \Phi\left(\frac{x-2}{2}\right)$. Thus,

$$\begin{aligned} F_X(3) &= \Phi\left(\frac{3-2}{2}\right) = \Phi(0.5) \approx 69.15\% \\ F_X(0) &= \Phi\left(\frac{0-2}{2}\right) = \Phi(-1) \approx 15.87\% \\ F_X(2) &= \Phi\left(\frac{2-2}{2}\right) = \Phi(0) = 0.5. \end{aligned}$$

□

5.1.3 Continuous Case

If X is a discrete rv, then $Y = g(X)$ is also a discrete rv. If X is a continuous rv, then we must put additional restrictions on the function $g(\cdot)$ in order to have $Y = g(X)$ be a continuous rv. We state the following theorem without proof:

Theorem 5.1.1. If X is a continuous rv and the function $g(\cdot)$ is differentiable at every real number x and further, $g'(x) \neq 0$ except for a finite number of values of x and $g'(x)$ changes sign only a finite number of times for $x \in \mathbb{R}$, then $Y = g(X)$ is a continuous rv. These conditions are sufficient but not necessary.

Now we assume that the hypotheses of theorem 5.1.1 are satisfied. Given a continuous random variable X, finding the pdf of a 1-1 function $Y = g(X)$ follows a straightforward process:

1. Find

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \le y) \\ &= \mathrm{P}(g(X) \le y) \\ &= \begin{Bmatrix} \mathrm{P}(X \le g^{-1}(y)) & g(\cdot) \text{ is increasing} \\ \mathrm{P}(X \ge g^{-1}(y)) & g(\cdot) \text{ is decreasing} \end{Bmatrix} \\ &= \begin{Bmatrix} \mathrm{P}(X \le g^{-1}(y)) & g(\cdot) \text{ is increasing} \\ 1 - \mathrm{P}(X < g^{-1}(y)) & g(\cdot) \text{ is decreasing} \end{Bmatrix} \\ &= \begin{Bmatrix} F_X(g^{-1}(y)) & g(\cdot) \text{ is increasing} \\ 1 - F_X(g^{-1}(y)) & g(\cdot) \text{ is decreasing} \end{Bmatrix}. \end{aligned}$$

2. Find $f_Y(y) = \frac{dF_Y(y)}{dy} = a_g F_X'(g^{-1}(y)) \frac{dg^{-1}(y)}{dy}$ (the chain rule) $= a_g f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy}$, where $a_g = \begin{Bmatrix} 1 & g(\cdot) \text{ increasing} \\ -1 & g(\cdot) \text{ decreasing} \end{Bmatrix}$.

In other words, if $Y = g(X)$ and $g(\cdot)$ is a 1-1 function,

$$\boxed{f_Y(y) = a_g f_X(g^{-1}(y)) \cdot \frac{dg^{-1}(y)}{dy}, \quad a_g = \begin{cases} 1 & g(\cdot) \text{ increasing} \\ -1 & g(\cdot) \text{ decreasing} \end{cases}.} \tag{5.1.1}$$

Furthermore, if $g(\cdot)$ is *not* 1-1, break it up into intervals that are 1-1. For instance, for $g(X) = X^2$, it is 1-1 on $(-\infty, 0]$ and on $(0, \infty)$, so that

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(X^2 \leq y) = \mathrm{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = \mathrm{P}(X \leq \sqrt{y}) - \mathrm{P}(X < -\sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}). \tag{5.1.2}$$

so that we can use the chain rule and get

$$f_Y(y) = f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} + f_X(-\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} \tag{5.1.3}$$

A good way to think about these problems is to remember the two steps above (i.e., find $F_Y(y)$, then take the derivative), so that you're prepared even if you don't remember (5.1.1) and/or the assumptions behind it.

Example 5.1.6. Let $X \sim N(\mu, \sigma^2)$, so that

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad x, \mu \in (-\infty, \infty); \sigma \in (0, \infty).$$

Let $Y = g(X) = \frac{X-\mu}{\sigma} \;\Rightarrow\; g^{-1}(y) = \sigma y + \mu$ and $g(\cdot)$ is increasing, so that by (5.1.1),

$$f_Y(y) = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} = \sigma f_X(\sigma y + \mu) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2},$$

so that $Y \sim N(0, 1)$. □

Example 5.1.7. Let X denote the molar concentration of a given compound. Let Y denote the absorbance at a given wave length:

$$Y = aX + b \equiv g(X) \quad \Rightarrow \quad g^{-1}(y) = \frac{y-b}{a}.$$

where $a > 0$ is a constant which depends on wave length of a compound. This compound is dissolved in a solvent. The constant b depends on the solvent. Assume $X \sim N(\mu, \sigma^2)$. Since $g(\cdot)$ is increasing, (5.1.1) tells us that

$$f_Y(y) = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} = \frac{1}{a} f_X\left(\frac{y-b}{a}\right) = \frac{1}{a\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\left(\frac{y-b}{a} - \mu\right)^2\right) = \frac{1}{a\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-b-\mu a)^2}{2a^2\sigma^2}\right),$$

so that $Y \sim N(b + \mu a, \sigma^2 a^2)$. □

Example 5.1.8. Assume that $X \sim \mathrm{U}(0, 1)$:

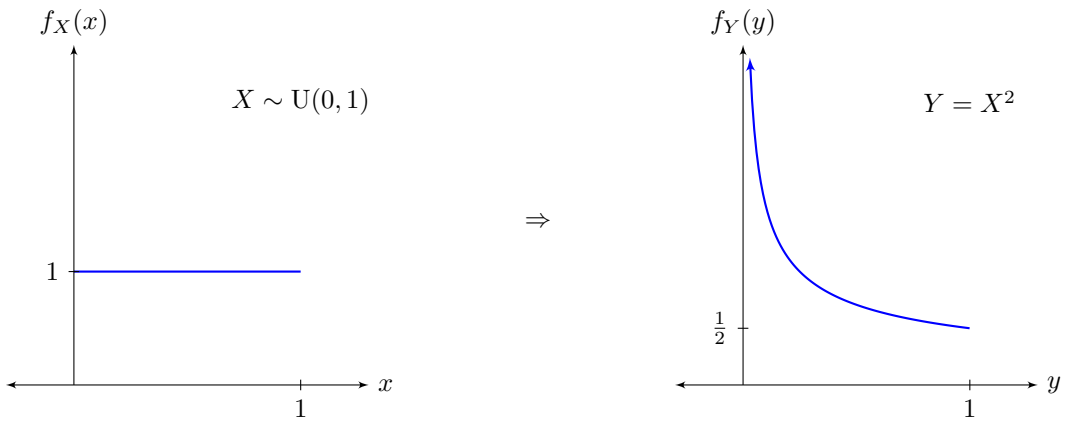
$$f_X(x) = 1, \quad x \in [0, 1].$$

Let $Y = X^2$, which is not 1-1, so that we can't use (5.1.1) [That is, since $g(\cdot)$ is not 1-1, it doesn't have an inverse!]. First we realize that the range of Y is from 0 to 1. Thus, for $0 \leq y \leq 1$, we first find $F_Y(y)$ and try to break it up:

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(X^2 \leq y) = \mathrm{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = \mathrm{P}(X \leq \sqrt{y}) - \mathrm{P}(X < -\sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

This is exactly the same equation as (5.1.2) – but we re-write it here because the process is the same for any $g(\cdot)$ that is not 1-1. Using (5.1.3), we have

$$f_Y(y) = f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} + f_X(-\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} = \frac{1}{2\sqrt{y}}[1 + 0] = \frac{1}{2\sqrt{y}}, \quad y \in [0, 1].$$ □



Example 5.1.9. Assume that the rv X has the pdf

$$f_X(x) = \frac{\lambda}{2} e^{-\lambda|x|}, \quad x \in \mathbb{R}.$$

Let $Y = X^2$. Then the range of Y is $[0, \infty)$, so that we have by (5.1.3) that

$$f_Y(y) = f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} + f_X(-\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} = \frac{\lambda e^{-\lambda\sqrt{y}}}{4\sqrt{y}} + \frac{\lambda e^{-\lambda\sqrt{y}}}{4\sqrt{y}} = \frac{\lambda e^{-\lambda\sqrt{y}}}{2\sqrt{y}}, \quad y \in (0, \infty).$$

Note that we started with the range $y \in [0, \infty)$, but then had to remove the point $y = 0$ to avoid division be zero.□

Example 5.1.10 (Random Sine Wave). Let $Y = a\sin(X)$. The amplitude a is a known positive constant. The angle X is a continuous r.v. whose pdf is

$$f_X(x) = \frac{1}{\pi}, \quad x \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right].$$

Y is a continuous rv, since $g'(x) = a\cos(x) > 0 \ \ \forall x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Therefore, we can use Theorem 5.1.1. Note that on this range of X, $g(X) = a\sin(X)$ is a 1-1, increasing function (with the range of $[-a, a]$), so that we can use (5.1.1). Therefore, $g^{-1}(y) = \sin^{-1}(\frac{y}{a})$, and

$$f_Y(y) = f_X(g^{-1}(y))\frac{dg^{-1}(y)}{dy} = \frac{1}{\pi a\sqrt{1 - y^2/a^2}}, \quad y \in [-a, a]$$

where we make use of the fact that $\frac{d}{dx}\sin^{-1}(u) = \frac{1}{\sqrt{1-u^2}}\frac{du}{dx}$.

If we don't remember (5.1.1), then we note that $-a \le y \le a$ when $-\frac{\pi}{2} \le x \le \frac{\pi}{2}$. For $|y| \le a$, $F_Y(y)$ is given by

$$F_Y(y) = \mathrm{P}_Y(Y \le y) = \mathrm{P}_X(a\sin(X) \le y) = \mathrm{P}_X\left(\sin(X) \le \frac{y}{a}\right) = \mathrm{P}_X\left(X \le \sin^{-1}\left(\frac{y}{a}\right)\right) = F_X\left(\sin^{-1}\left(\frac{y}{a}\right)\right).$$

Thus, we take the derivative of the above to get

$$f_Y(y) = f_X\left(\sin^{-1}\left(\frac{y}{a}\right)\right) \cdot \frac{1}{a}\left(1 - \frac{y^2}{a^2}\right)^{-1/2} = \left(\frac{1}{\pi a}\right)\left(\frac{1}{\sqrt{1 - y^2/a^2}}\right), \quad |y| \le a.$$

A third way to do this is to note that

$$F_Y(y) = \int_{-\pi/2}^{\sin^{-1}(y/a)} \frac{1}{\pi}\, dx = \frac{1}{\pi}\left(\sin^{-1}(y/a) + \pi/2\right).$$

Now take the derivative, and recall that $\frac{d}{dx}\sin^{-1}(u) = (1-u^2)^{-1/2}\frac{du}{dx}$. Thus,

$$f_Y(y) = \frac{1}{\pi a}\frac{1}{\sqrt{1-y^2/a^2}}, \quad |y| \le a. \qquad \square$$

5.1.4 Discrete Case

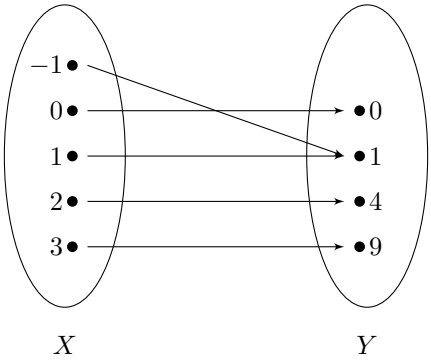
Let $Y = g(X) = X^2$. Then

$$p_Y(y) = \mathrm{P}_Y(Y=y) = \mathrm{P}_X(X^2=y) = \mathrm{P}_X(X=\sqrt{y} \text{ or } X=-\sqrt{y}) = p_X(\sqrt{y}) + p_X(-\sqrt{y}), \quad y \ge 0.$$

Example 5.1.11. Suppose X is discrete with pmf

$$p_X(x) = \frac{x+2}{15}, \quad x \in \{-1,0,1,2,3\}.$$

Then for $Y = X^2$, Y has values of $\{0,1,4,9\}$:

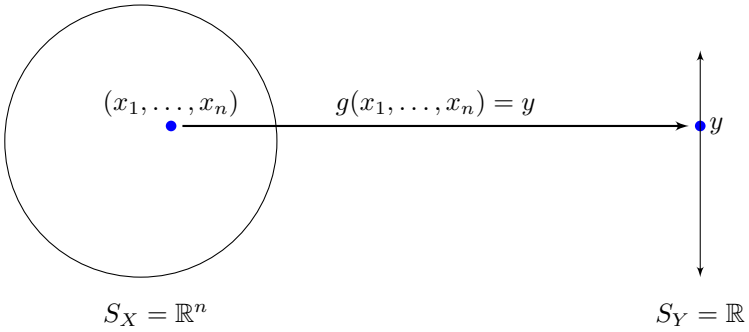


That is,

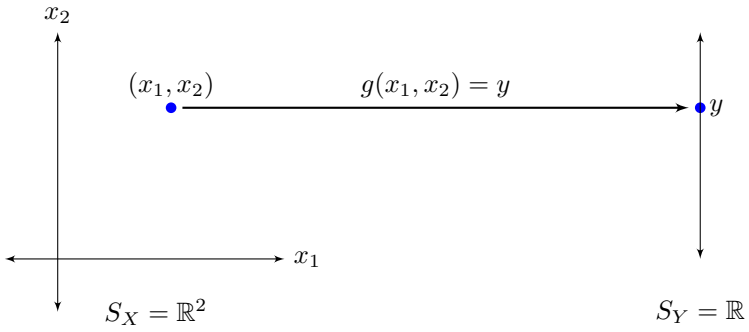
$$\begin{aligned}
p_Y(0) &= p_X(0) = \frac{2}{15} \approx 13.33\%, \\
p_Y(1) &= p_X(\sqrt{1}) + p_X(-\sqrt{1}) = \frac{1}{15} + \frac{3}{15} = \frac{4}{15} \approx 26.67\%, \\
p_Y(4) &= p_X(\sqrt{4}) = p_X(2) = \frac{4}{15} \approx 26.67\%, \\
p_Y(9) &= p_X(\sqrt{9}) = p_X(3) = \frac{5}{15} = 33.33\%.
\end{aligned} \qquad \square$$

5.2 Functions of Many Random Variables

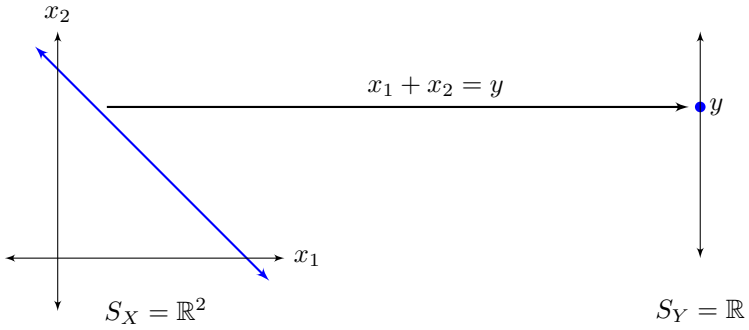
Let the rv Y be denoted by $Y = g(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ are jointly distributed random variables.



We will first restrict our attention to the two-dimensional case:



Consider the transformation $Y = X_1 + X_2 = g(X_1, X_2)$:



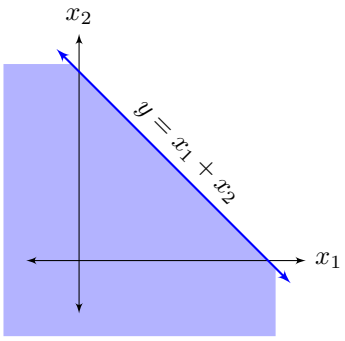
As we see from this concrete example, it often isn't *one* point (x_1, x_2) that is mapped to a value of y, but rather a *set of points*. That is,

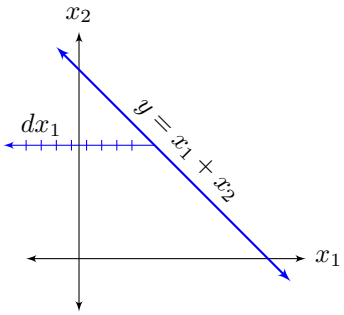
$$\begin{aligned}
&\text{The set } \{(x_1, x_2) : \ x_1 + x_2 = 6\} \text{ is mapped to } y = 6, \\
&\text{The set } \{(x_1, x_2) : \ x_1 + x_2 = 8\} \text{ is mapped to } y = 8, \\
&\text{The set } \{(x_1, x_2) : \ x_1 + x_2 = -3\} \text{ is mapped to } y = -3, \\
&\qquad \text{etc.}
\end{aligned}$$

5.2.1 Continuous Case: $Y = X_1 + X_2$

Let X_1 and X_2 be jointly continuous rvs, where $S = \mathbb{R} \times \mathbb{R} = \mathbb{R}^2$. Let the joint pdf for X_1, X_2 be $f_{\boldsymbol{X}}(x_1, x_2)$, and let $Y = X_1 + X_2$. Thus,

$$\begin{aligned} F_Y(y) &= \mathrm{P}_Y(Y \le y) \\ &= \mathrm{P}_{X_1,X_2}(X_1 + X_2 \le y) \\ &= \mathrm{P}_{X_1,X_2}(\{x_1, x_2 :\ x_1 + x_2 \le y\}) \\ &= \iint\limits_{\{(x_1,x_2):\ x_1+x_2\le y\}} f_{\boldsymbol{X}}(x_1, x_2)\ dx_1 dx_2 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{y-x_2} f_{\boldsymbol{X}}(x_1, x_2)\ dx_1 dx_2 \end{aligned}$$



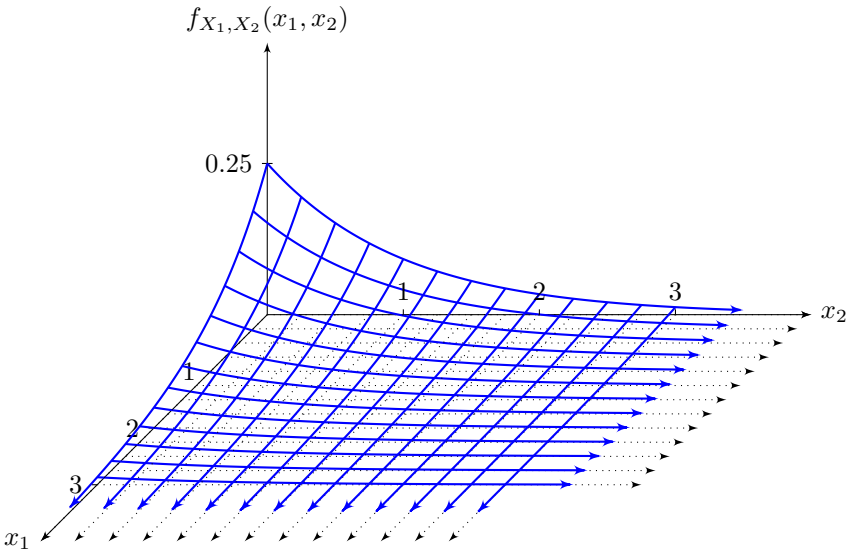


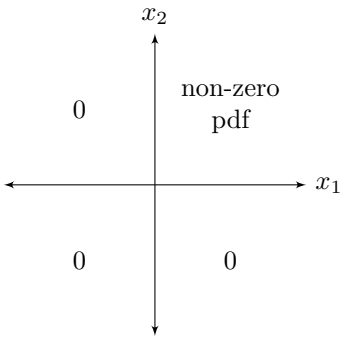
Example 5.2.1. Let X_1 and X_2 be independent rvs with the pdf

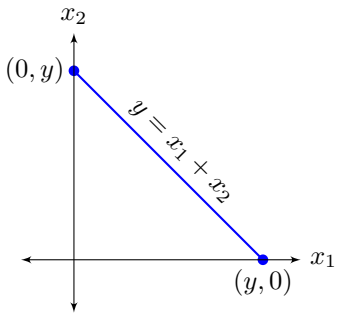
$$f_X(x) = \frac{1}{2} e^{-x/2}, \quad x \ge 0.$$

Since X_1 and X_2 are independent, we have

$$f_{\boldsymbol{X}}(x_1, x_2) = f_X(x_1) \cdot f_X(x_2) = \frac{1}{2} e^{-x_1/2} \cdot \frac{1}{2} e^{-x_2/2}, \quad x_1, x_2 \ge 0.$$







For a practical application of this problem, let X_i denote the life length of a given piece of equipment in months. We have a spare on hand so that when the equipment fails for the first time, we replace it. The average life length for each of them is 2 months. Thus, X_1 is the life length of the original piece, and X_2 is the life length of the spare. We want to determine the life length of $Y = X_1 + X_2$ (which will always be ≥ 0:

$$\begin{aligned}
F_Y(y) &= \mathrm{P}_Y(Y \leq y) \\
&= \mathrm{P}_{X_1,X_2}(X_1 + X_2 \leq y) \\
&= \mathrm{P}_{X_1,X_2}(\{(x_1, x_2) : \ x_1 + x_2 \leq y\}) \\
&= \int_0^y \int_0^{y-x_2} \frac{1}{2} e^{-x_1/2} \cdot \frac{1}{2} e^{-x_2/2} \, dx_1 dx_2 \\
&= \int_0^y \frac{1}{2} e^{-x_2/2} \left[\int_0^{y-x_2} \frac{1}{2} e^{-x_1/2} \, dx_1 \right] \, dx_2 \\
&= \int_0^y \frac{1}{2} e^{-x_2/2} \left[-e^{-x_1/2} \right]_0^{y-x_2} \, dx_2 \\
&= \int_0^y \frac{1}{2} e^{-x_2/2} \left(1 - e^{(y-x_2)/2} \right) \, dx_2 \\
&= \int_0^y \frac{1}{2} \left(e^{-y/2} + e^{-x_2/2} \right) \, dx_2 \\
&= -\frac{y}{2} e^{-y/2} - e^{-y/2} + 1
\end{aligned}$$

Therefore,

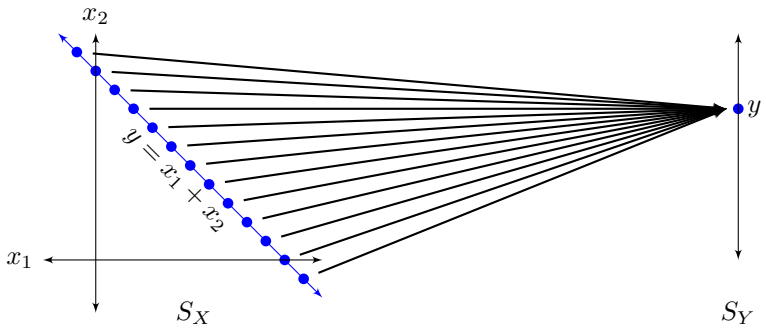
$$\begin{aligned}
f_Y(y) &= F_Y'(y) \\
&= -\frac{y}{2} \left(-\frac{1}{2} e^{-y/2} \right) + e^{-y/2} \left(-\frac{1}{2} \right) + \frac{1}{2} e^{-y/2} \\
&= \frac{y}{4} e^{-y/2}, \quad y \geq 0.
\end{aligned}$$

□

5.2.2 Discrete Case: $Y = X_1 + X_2$

Let X_1 and X_2 be discrete random variables with joint pmf $p_{\boldsymbol{X}}(x_1, x_2)$. Let $Y = X_1 + X_2$. The pmf for Y is

$$\begin{aligned} p_Y(y) &= \mathrm{P}_Y(Y = y) \\ &= \mathrm{P}_{X_1,X_2}(X_1 + X_2 = y) \\ &= \mathrm{P}_{X_1,X_2}(\{(x_1, x_2) : \ x_1 + x_2 = y\}) \\ &= \sum_{\substack{\text{all } x_1 \ni \\ p_{\boldsymbol{X}}(x_1, y-x_1)>0}} p_{\boldsymbol{X}}(x_1, y - x_1). \end{aligned}$$



Example 5.2.2. Let the discrete rvs X_1 and X_2 be independent with pmfs

$$p_{X_i}(x_i) = \frac{1}{6}, \quad x_i \in \{1, 2, 3, 4, 5, 6\}.$$

Since X_1 and X_2 are independent,

$$p_{\boldsymbol{X}}(x_1, x_2) = p_{X_1}(x_1)p_{X_2}(x_2) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}, \quad x_1, x_2 \in \{1, 2, 3, 4, 5, 6\}.$$

Let $Y = X_1 + X_2$. Then Y has possible values in $\{2, 3, \ldots, 11, 12\}$, and

$$p_Y(y) = \sum_{\substack{\text{all } x_1 \ni \\ p_{\boldsymbol{X}}(x_1, y-x_1)>0}} p_{\boldsymbol{X}}(x_1, y - x_1), \quad y \in \{2, 3, \ldots, 12\}.$$

For example,

$$p_Y(5) = \sum_{\substack{\text{all } x_1 \ni \\ p_{\boldsymbol{X}}(x_1, 5-x_1)>0}} p_{\boldsymbol{X}}(x_1, 5 - x_1) = p_{\boldsymbol{X}}(1, 4) + p_{\boldsymbol{X}}(2, 3) + p_{\boldsymbol{X}}(3, 2) + p_{\boldsymbol{X}}(4, 1) = \frac{4}{36},$$

$$p_Y(7) = \sum_{\substack{\text{all } x_1 \ni \\ p_{\boldsymbol{X}}(x_1, 7-x_1)>0}} p_{\boldsymbol{X}}(x_1, 7 - x_1) = p_{\boldsymbol{X}}(1, 6) + p_{\boldsymbol{X}}(2, 5) + p_{\boldsymbol{X}}(3, 4) + p_{\boldsymbol{X}}(4, 3) + p_{\boldsymbol{X}}(5, 2) + p_{\boldsymbol{X}}(6, 1) = \frac{6}{36}. \quad \square$$

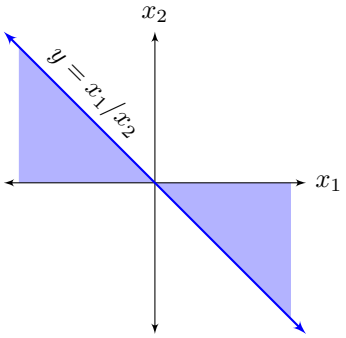
Now let's consider a transformation which is not linear.

Example 5.2.3. Let X_1 and X_2 be independent $\mathrm{N}(0,1)$ rvs:

$$f_{\boldsymbol{X}}(x_1,x_2) = \frac{1}{\sqrt{2\pi}}e^{-x_1/2}\cdot\frac{1}{\sqrt{2\pi}}e^{-x_2/2}, \quad x_1,x_2\in\mathbb{R}.$$

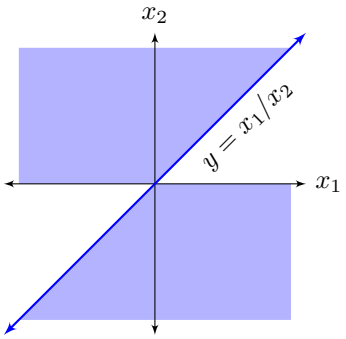
Let $Y = X_1/X_2 \quad\Rightarrow\quad Y\in\mathbb{R}$.

For $y<0$,



$$\begin{aligned}F_Y(y) &= \mathrm{P}_{\boldsymbol{X}}(X_1<yX_2 \text{ and } X_2>0) + \mathrm{P}_{\boldsymbol{X}}(X_1>yX_2 \text{ and } X_2<0)\\ &= \int_0^\infty\int_{-\infty}^{yx_2} f_{\boldsymbol{X}}(x_1,x_2)\,dx_1dx_2 + \int_{-\infty}^0\int_{yx_2}^\infty f_{\boldsymbol{X}}(x_1,x_2)\,dx_1dx_2\end{aligned}$$

For $y>0$,



$$\begin{aligned}F_Y(y) &= \mathrm{P}_{\boldsymbol{X}}(X_1<yX_2 \text{ and } X_2>0) + \mathrm{P}_{\boldsymbol{X}}(X_1>yX_2 \text{ and } X_2>0)\\ &= \int_0^\infty\int_{-\infty}^{yx_2} f_{\boldsymbol{X}}(x_1,x_2)\,dx_1dx_2 + \int_{-\infty}^0\int_{yx_2}^\infty f_{\boldsymbol{X}}(x_1,x_2)\,dx_1dx_2\end{aligned}$$

Since the expression for $F_Y(y)$ is the same for $y<0$ as for $y>0$, we differentiate this expression for $F_Y(y)$ to obtain $f_Y(y)\ \forall y\in\mathbb{R}$:[1]

$$\begin{aligned}f_Y(y) &= \int_0^\infty \frac{x_2}{2\pi}\exp\left(-\frac{(yx_2)^2}{2}\right)\cdot\exp\left(-\frac{x_2^2}{2}\right)\,dx_2 - \int_{-\infty}^0 \frac{x_2}{2\pi}\exp\left(-\frac{(yx_2)^2}{2}\right)\cdot\exp\left(-\frac{x_2^2}{2}\right)\,dx_2\\ &= \frac{1}{2\pi}\int_0^\infty x_2\exp\left(-\frac{x_2^2}{2}(y^2+1)\right)\,dx_2 - \frac{1}{2\pi}\int_{-\infty}^0 x_2\exp\left(-\frac{x_2^2}{2}(y^2+1)\right)\,dx_2\\ &= \frac{2}{2\pi}\int_0^\infty x_2\exp\left(-\frac{x_2^2}{2}(y^2+1)\right)\,dx_2\\ &= -\frac{1}{\pi(y^2+1)}\left[-\frac{x_2^2(y^2+1))}{2}\right]_0^\infty\\ &= \frac{1}{\pi(y^2+1)}, \quad y\in\mathbb{R}.\end{aligned}$$

This distribution is called the *Cauchy distribution*. □

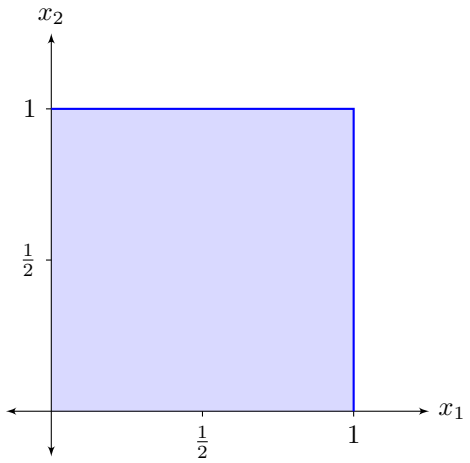
[1] Recall that

$$\frac{d}{dy}\int_a^b\int_{g(y)}^{h(y)} f_{\boldsymbol{X}}(x_1,x_2)\,dx_1dx_2 = \int_a^b\left[f_{\boldsymbol{X}}\left(h(y),x_2\right)\cdot h'(y)\right]\,dx_2 - \int_a^b\left[f_{\boldsymbol{X}}\left(g(y),x_2\right)\cdot g'(y)\right]\,dx_2.$$

Example 5.2.4. Let $X_1, X_2 \overset{iid}{\sim} \mathrm{U}(0,1)$. Thus, our joint pdf is

$$f_{X_1X_2}(x_1,x_2) = f_{X_1}(x_1)f_{X_2}(x_2) = 1_{[x_1\in[0,1]]}1_{[x_2\in[0,1]]} = 1 \text{ for } x_1,x_2 \in [0,1],$$

so that $0 \le X_1 \le 1$ and $0 \le X_2 \le 1$. If we make a graph of the points (x_1,x_2) that have positive (i.e., nonzero) probability, we see that it is a square from the origin to the point $(x_1,x_2) = (1,1)$. This is called our *region of support*:



Find the pdf of the functions given:

(a) $X_1 + X_2$.

We know that possible values of $Y = X_1 + X_2$ are between 0 and 2, inclusive. To find the pdf, we first find $F_Y(y)$ and then take its derivative. To start this process, we have

$$F_Y(y) = \mathrm{P}(Y \le y) = \mathrm{P}(X_1 + X_2 \le y) = \mathrm{P}(X_2 \le y - X_1).$$

If we graph this, out region is below the line $x_2 = y - x_1$. Figure 5.1 shows this region (in dark blue) for the case where $0 < y \le 1$ and where $1 < y < 2$. Here we see that these two different sets of values for y gives us two different regions. We will set up our double integrals from these graphs.

If $y \in [0,1]$,

$$\begin{aligned} F_Y(y) &= \mathrm{P}(X_2 \le y - X_1) \\ &= \int_0^y \int_0^{y-x_1} 1 \; dx_2 dx_1 \\ &= \int_0^y y - x_1 \; dx_1 \\ &= y^2 - \frac{y^2}{2} \\ &= \frac{y^2}{2}, \end{aligned}$$

so that $f_Y(y) = \frac{d}{dy}\left(\frac{y^2}{2}\right) = y$.

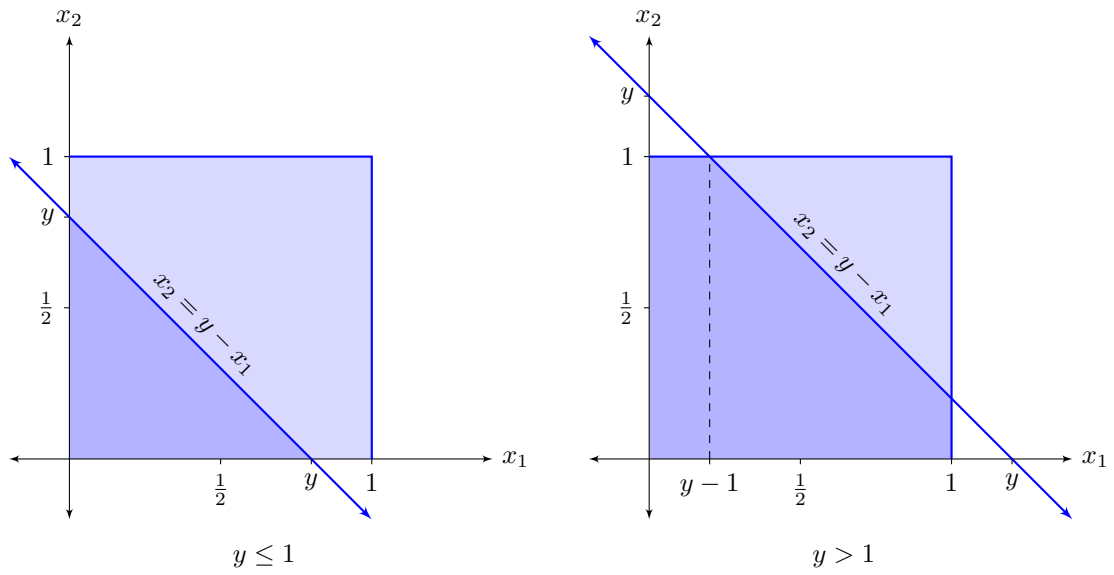


Figure 5.1: The area (in dark blue) representing $X_1 + X_2 \leq y$, for $y \leq 1$ and $y > 1$.

If $y \in (1, 2]$, we can see by the graph on the right of Figure 5.1 that

$$\begin{aligned} F_Y(y) &= \mathrm{P}(X_2 \leq y - X_1) \\ &= \int_0^{y-1} \int_0^1 1 \; dx_2 dx_1 + \int_{y-1}^1 \int_0^{y-x_1} 1 \; dx_2 dx_1 \qquad \text{partition on both sides of } x_1 = y - 1 \text{ as in Figure 5.1} \\ &= y - 1 + \int_{y-1}^1 y - x_1 \; dx_1 \\ &= y - 1 + y - \frac{1}{2} - y(y-1) - \frac{(y-1)^2}{2} \\ &= 2y - 1 - \frac{y^2}{2}, \end{aligned}$$

so that $f_Y(y) = \frac{d}{dy}\left(2y - 1 - \frac{y^2}{2}\right) = 2 - y$.

Note that for $y \in (1, 2]$, since we know that the integral over the region of support (the light blue square) is equal to 1, we could have also taken the integral of the upper right triangle and subtracted that from 1:

$$\begin{aligned}
F_Y(y) &= \mathrm{P}(X_2 \le y - X_1) \\
&= 1 - \mathrm{P}(X_2 > y - X_1) \\
&= 1 - \int_{y-1}^{1} \int_{y-x_1}^{1} 1 \; dx_2 dx_1 \\
&= 1 - \int_{y-1}^{1} 1 - (y - x_1) \; dx_1 \\
&= 1 - \int_{y-1}^{1} (1-y) + x_1 \; dx_1 \\
&= 1 - \left[(1-y)x_1 + \frac{x_1^2}{2} \right]_{y-1}^{1} \\
&= 1 - \left(1 - y + \frac{1}{2} - (1-y)(y-1) - \frac{(y-1)^2}{2} \right) \\
&= 2y - 1 - \frac{y^2}{2}.
\end{aligned}$$

Altogether,

$$f_Y(y) = \begin{cases} y & y \in [0,1] \\ 2-y & y \in (1,2] \\ 0 & \text{otherwise} \end{cases},$$

which integrates to 1 (not shown here).

(b) $X_1 - X_2$.

The possible values of $Y = X_1 - X_2$ are between -1 and 1, inclusive. Once again, to find the pdf, we first find $F_Y(y)$ and then take its derivative. To start this process, we have

$$F_Y(y) = \mathrm{P}(Y \le y) = \mathrm{P}(X_1 - X_2 \le y) = \mathrm{P}(-X_2 \le y - X_1) = \mathrm{P}(X_2 \ge X_1 - y).$$

Note that the direction of the inequality is now switched because we multiplied by a negative number. If we graph this, out region is now *above* the line $x_2 = x_1 - y$. Figure 5.2 shows this region (in dark blue) for the case where $-1 < y \le 0$ and where $0 < y < 1$. Here we see that these two different sets of values for y gives us two different regions. As before, we will set up our double integrals from these graphs.

The possible values of $Y = X_1 - X_2$ are between -1 and 1, inclusive. If $y \le 0$,

$$\begin{aligned}
F_Y(y) &= \mathrm{P}(X_2 \ge X_1 - y) \\
&= \int_{0}^{1+y} \int_{x_1-y}^{1} 1 \; dx_2 dx_1 \\
&= \int_{0}^{1+y} 1 - (x_1 - y) \; dx_1 \\
&= \int_{0}^{1+y} (1+y) - x_1 \; dx_1 \\
&= \left[(1+y)x_1 - \frac{x_1^2}{2} \right]_{0}^{1+y} \\
&= \frac{(1+y)^2}{2}.
\end{aligned}$$

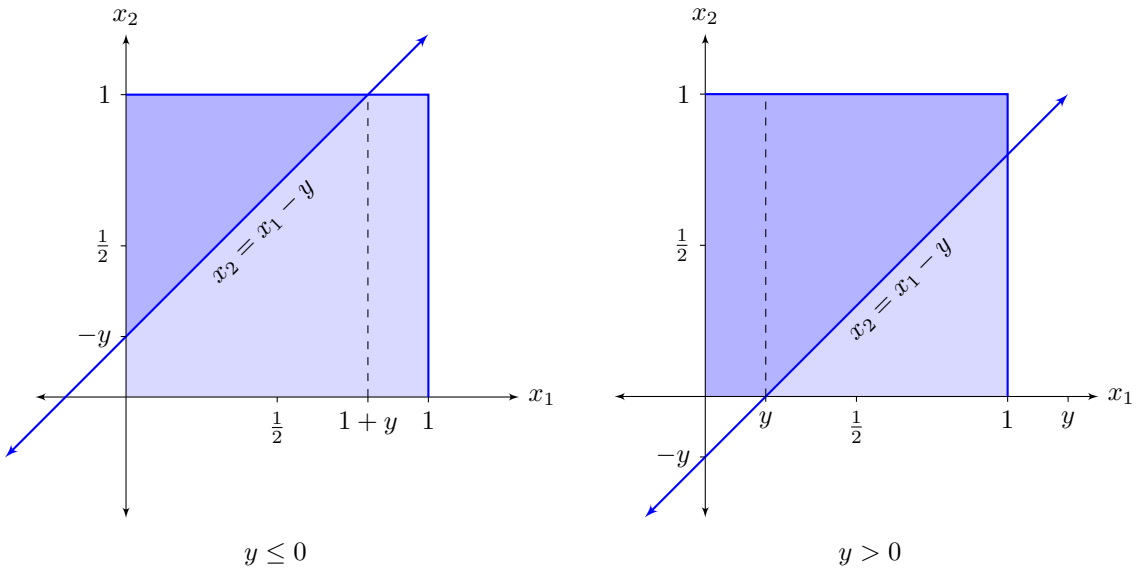


Figure 5.2: The area (in dark blue) representing $X_2 \geq X_1 - y$, for $y \leq 0$ and $y > 0$.

so that $f_Y(y) = \frac{d}{dy}\left(\frac{(1+y)^2}{2}\right) = 1 + y$.

If $y > 0$, we can see from the right side of Figure 5.2 that

$$\begin{aligned}
F_Y(y) &= \mathrm{P}(X_2 \geq X_1 - y) \\
&= \int_0^y \int_0^1 dx_2 dx_1 + \int_y^1 \int_{x_1 - y}^1 dx_2 dx_1 \\
&= y + \int_y^1 1 - (x_1 - y)\, dx_1 \\
&= y + \int_y^1 (1 + y) - x_1\, dx_1 \\
&= y + \left[(1+y)x_1 - \frac{x_1^2}{2}\right]_y^1 \\
&= y + (1+y) - \frac{1}{2} - \left((1+y)y - \frac{y^2}{2}\right) \\
&= 2y + 1 - \frac{1}{2} - (y + y^2 - \frac{y^2}{2}) \\
&= y + \frac{1}{2} - \frac{y^2}{2},
\end{aligned}$$

so that $f_Y(y) = \frac{d}{dy}\left(y + \frac{1}{2} - \frac{y^2}{2}\right) = 1 - y$.

As in (a), we could have also taken the integral of the lower right triangle and subtracted it from 1:

$$\begin{aligned}
F_Y(y) &= \mathrm{P}(X_2 \geq X_1 - y) \\
&= 1 - \mathrm{P}(X_2 < X_1 - y) \\
&= 1 - \int_y^1 \int_0^{x_1 - y} 1 \; dx_2 dx_1 \\
&= 1 - \int_y^1 x_1 - y \; dx_1 \\
&= 1 - \left[\frac{x_1^2}{2} - y x_1 \right]_y^1 \\
&= 1 - \left(\frac{1}{2} - y - \left(\frac{y^2}{2} - y^2 \right) \right) \\
&= y + \frac{1}{2} - \frac{y^2}{2}.
\end{aligned}$$

Altogether,

$$f_Y(y) = \begin{cases} 1 + y & y \in [-1, 0] \\ 1 - y & y \in (0, 1] \\ 0 & \text{otherwise} \end{cases},$$

which integrates to 1.

(c) $\min(X_1, X_2)$.

The possible values of $Y = \min(X_1, X_2)$ are between 0 and 1, inclusive. We don't need a graph for this one. For $y \in [0, 1]$,

$$\begin{aligned}
F_Y(y) &= \mathrm{P}(Y \leq y) \\
&= \mathrm{P}(\min(X_1, X_2) \leq y) \\
&= \mathrm{P}(X_1 \leq y, \; X_2 \leq y) + \mathrm{P}(X_1 \leq y, \; X_2 \geq y) + \mathrm{P}(X_1 \geq y, \; X_2 \leq y) && \text{mutually exclusive events} \\
&= \mathrm{P}(X_1 \leq y)\mathrm{P}(X_2 \leq y) + \mathrm{P}(X_1 \leq y)\mathrm{P}(X_2 \geq y) + \mathrm{P}(X_1 \geq y)\mathrm{P}(X_2 \leq y) && \text{independence} \\
&= y^2 + y(1 - y) + (1 - y)y && \text{since } F_X(y) = y \text{ for } X \sim U(0, 1) \\
&= 2y - y^2.
\end{aligned}$$

so that $f_Y(y) = \frac{d}{dy}\left(2y - y^2\right) = 2 - 2y$, which can be shown to integrate to 1 over $y \in [0, 1]$.

Alternatively, an easier way to compute $F_Y(y)$ is the following:

$$\begin{aligned}
F_Y(y) &= \mathrm{P}(Y \leq y) \\
&= \mathrm{P}(\min(X_1, X_2) \leq y) \\
&= 1 - \mathrm{P}(\min(X_1, X_2) > y) \\
&= 1 - \mathrm{P}(X_1 > y, X_2 > y) \\
&= 1 - \mathrm{P}(X_1 > y)\mathrm{P}(X_2 > y) && \text{independence} \\
&= 1 - (1 - y)^2 && \text{since } 1 - F_X(y) = 1 - y \text{ for } X \sim U(0, 1) \\
&= 2y - y^2.
\end{aligned}$$

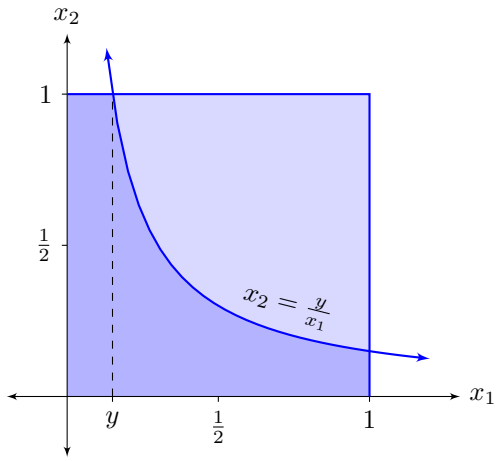


Figure 5.3: The area (in dark blue) representing $X_2 \leq \frac{y}{X_1}$.

(d) $\max(X_1, X_2)$.

The possible values of $Y = \max(X_1, X_2)$ are between 0 and 1, inclusive. We don't need a graph for this one. For $y \in [0, 1]$,

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \leq y) \\ &= \mathrm{P}(\max(X_1, X_2) \leq y) \\ &= \mathrm{P}(X_1 \leq y,\ X_2 \leq y) \\ &= \mathrm{P}(X_1 \leq y)\mathrm{P}(X_2 \leq y) && \text{independence} \\ &= y^2. && \text{since } F_X(y) = y \text{ for } X \sim U(0,1) \end{aligned}$$

so that $f_Y(y) = \frac{d}{dy}\left(y^2\right) = 2y$, which can be shown to integrate to 1 over $y \in [0, 1]$.

(e) $X_1 X_2$.

The possible values of $Y = X_1 X_2$ are between 0 and 1, inclusive. For $y \in [0, 1]$,

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \leq y) \\ &= \mathrm{P}(X_1 X_2 \leq y) \\ &= \mathrm{P}(X_2 \leq \frac{y}{X_1}) \\ &= \int_0^y \int_0^1 dx_2 dx_1 + \int_y^1 \int_0^{y/x_1} dx_2 dx_1 && \text{partition on both sides of } x_1 = y, \text{ as in Figure 5.3} \\ &= y + \int_y^1 \frac{y}{x_1}\, dx_1 \\ &= y + y\left[\ln x_1\right]_y^1 \\ &= y - y \ln y \end{aligned}$$

□

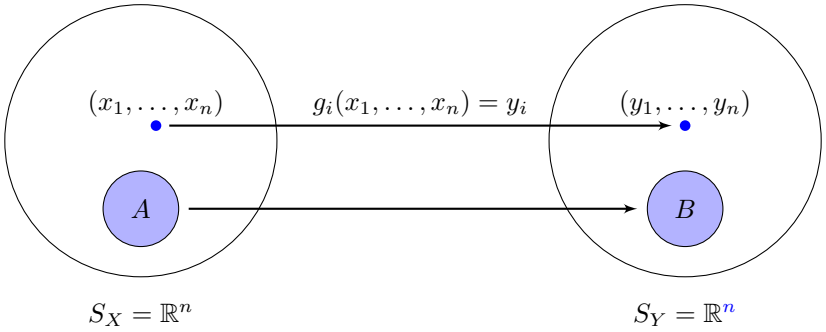
so that $f_Y(y) = \frac{d}{dy}(y - y \ln y) = -\ln y$, which is positive on the domain of y and which can be shown to integrate to 1 over $y \in (0, 1]$. Note that we should delete the point $y = 0$ to avoid division by zero.

5.3 Transformations from n-space to n-space

Let the n rvs $Y_1, \ldots, Y_n$ be denoted by

$$\begin{aligned} Y_1 &= g_1(X_1, \ldots, X_n) = g_1(\boldsymbol{X}), \\ Y_2 &= g_2(X_1, \ldots, X_n) = g_2(\boldsymbol{X}), \\ &\vdots \\ Y_n &= g_n(X_1, \ldots, X_n) = g_n(\boldsymbol{X}), \end{aligned}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, $\boldsymbol{X} = (X_1, \ldots, X_n)$, and where $X_1, \ldots, X_n$ are jointly distributed random variables.



We define probability in the $\boldsymbol{Y}$ space by

$$\mathrm{P}_{\boldsymbol{Y}}((Y_1, \ldots, Y_n) \in B) = \mathrm{P}_{\boldsymbol{X}}(g^{-1}(B)) = \mathrm{P}_{\boldsymbol{X}}(A) = \mathrm{P}_{\boldsymbol{X}}((X_1, \ldots, X_n) \in A)).$$

We need to determine sufficient conditions for $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ to have a joint continuous distribution and then to determine the joint pdf of $Y_1, \ldots, Y_n$.

Example 5.3.1. Transformations from 2-space to 2-space:

$$\begin{aligned} Y_1 &= X_1 + X_2 \\ Y_2 &= X_1 \end{aligned} \quad \text{or} \quad \begin{aligned} Y_1 &= X_1 X_2 \\ Y_2 &= X_1 - X_2 \end{aligned} \quad \text{or} \quad \begin{aligned} Y_1 &= X_1 / X_2 \\ Y_2 &= X_2 \end{aligned} \quad \text{or} \quad \begin{aligned} Y_1 &= X_1 \sin(X_2) \\ Y_2 &= X_1 \end{aligned} \quad . \ \square$$

Theorem 5.3.1. Let $X_1, \ldots, X_n$ be n random variables with a joint continuous distribution. Let

$$\begin{aligned} Y_1 &= g_1(X_1, \ldots, X_n) = g_1(\boldsymbol{X}), \\ Y_2 &= g_2(X_1, \ldots, X_n) = g_2(\boldsymbol{X}), \\ &\vdots \\ Y_n &= g_n(X_1, \ldots, X_n) = g_n(\boldsymbol{X}), \end{aligned}$$

be a mapping of $\mathbb{R}^n$ to $\mathbb{R}^n$ with these properties:

1. The mapping is 1-1.
2. The mapping and its inverse $\{x_i = h_i(y_1, \ldots, y_n) : \ i = 1, \ldots, n\}$ are continuous.
3. The n^2 partial derivatives $\frac{\partial x_i}{\partial y_j}$ for $i, j \in \{1, \ldots, n\}$ exist and are continuous.
4. The Jacobian

$$J = \frac{\partial(x_1, \ldots, x_n)}{\partial(y_1, \ldots, y_n)} \equiv \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix} \neq 0.$$

Then $Y_1, \ldots, Y_n$ have a joint continuous distribution with the joint pdf given by

$$f_{\boldsymbol{Y}}(y_1, \ldots, y_n) = f_{\boldsymbol{X}}(\ h_1(y_1, \ldots, y_n),\ \ldots,\ h_n(y_1, \ldots, y_n)\ ) \cdot |J|.$$

Proof: Advanced multivariate calculus text. ■

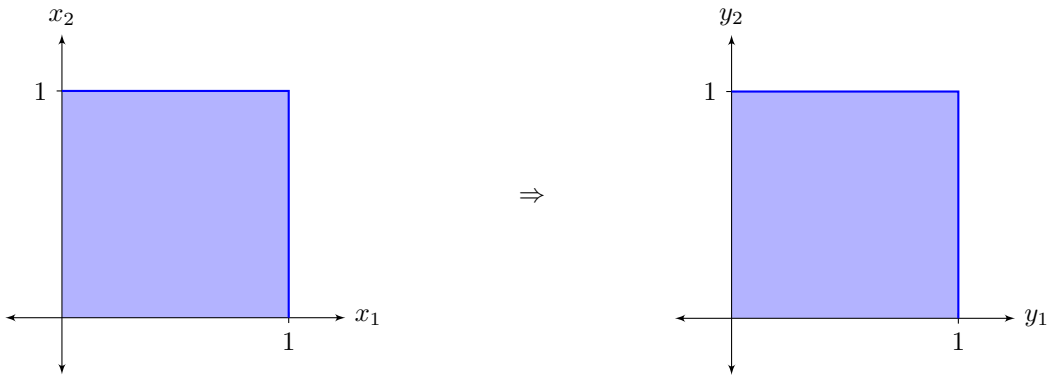
Example 5.3.2. Suppose X_1 and X_2 are jointly distributed with the pmf

$$f_{\boldsymbol{X}}(x_1, x_2) = 4x_1x_2, \quad x_1, x_2 \in (0, 1).$$

We would like to find the joint pdf of X_1^2 and X_2^2. First we need to find their inverse functions:

$$\begin{aligned} Y_1 &= X_1^2 = g_1(X_1, X_2) \\ Y_2 &= X_2^2 = g_2(X_1, X_2) \end{aligned} \quad \Rightarrow \quad \begin{aligned} X_1 &= \sqrt{Y_1} = h_1(Y_1, Y_2) \\ X_2 &= \sqrt{Y_2} = h_2(Y_1, Y_2) \end{aligned} .$$

Note that the range of (Y_1, Y_2) is the same as for (X_1, X_2): $(0, 1) \times (0, 1)$.



The Jacobian is

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} \frac{1}{2\sqrt{y_1}} & 0 \\ 0 & \frac{1}{2\sqrt{y_2}} \end{vmatrix} = \frac{1}{4\sqrt{y_1 y_2}} \neq 0.$$

Therefore,

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(\sqrt{y_1}, \sqrt{y_2}) \cdot |J| = 4\sqrt{y_1}\sqrt{y_2} \cdot \frac{1}{4\sqrt{y_1 y_2}} = 1, \quad y_1, y_2 \in (0, 1).$$

The marginal for Y_1 is

$$f_{Y_1}(y_1) = \int_0^1 f_{\boldsymbol{Y}}(y_1, y_2)\, dy_2 = \int_0^1 1\, dy_2 = 1, \quad y_1 \in (0, 1).$$

Similarly, the marginal for Y_2 is

$$f_{Y_2}(y_2) = 1, \quad y_2 \in (0, 1).$$

Note that Y_1 and Y_2 are independent, since

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1) \cdot f_{Y_2}(y_2), \quad y_1, y_2 \in (0, 1).$$ □

Example 5.3.3. Suppose X_1 and X_2 are jointly distributed with the pmf

$$f_{\boldsymbol{X}}(x_1, x_2) = 3x_1, \quad 0 < x_2 < x_1 < 1.$$

We want to find the pdf for $Y_1 = X_1 - X_2$. Again we first need to find their inverse functions:

$$\begin{aligned} Y_1 &= X_1 - X_2 = g_1(X_1, X_2) \\ Y_2 &= X_2 = g_2(X_1, X_2) \end{aligned} \quad \Rightarrow \quad \begin{aligned} X_1 &= Y_1 + Y_2 = h_1(Y_1, Y_2) \\ X_2 &= Y_2 = h_2(Y_1, Y_2) \end{aligned} .$$

The Jacobian is

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = 1 \neq 0.$$

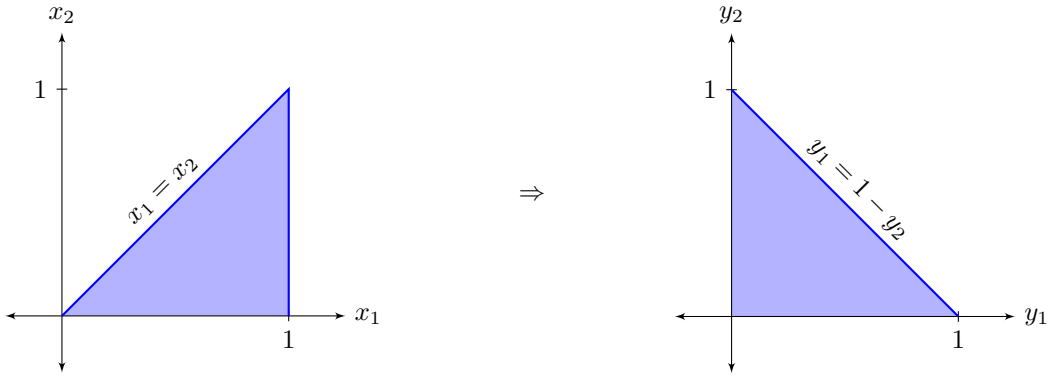
For the range of (Y_1, Y_2), note that

$$0 < x_2 < x_1 < 1 \quad \Leftrightarrow \quad 0 < y_2 < y_1 + y_2 < 1 \quad \Rightarrow \quad 0 < y_1 < 1 - y_2.$$

and

$$0 < x_2 < 1 \quad \Rightarrow \quad 0 < y_2 < 1.$$

Therefore, our regions of integration are



For the region indicated,

$$\begin{aligned} f_{\boldsymbol{Y}}(y_1, y_2) &= f_{\boldsymbol{X}}(y_1 + y_2, y_2) \cdot 1 \\ &= 3(y_1 + y_2), \end{aligned}$$

$$\begin{aligned} f_{Y_1}(y_1) &= \int_{S_{Y_2}} f_{Y_1, Y_2}(y_1, y_2)\, dy_2 \\ &= \int_0^{1-y_1} 3(y_1 + y_2)\, dy_2 \\ &= \left[3\left(y_1 y_2 + \frac{y_2^2}{2} \right) \right]_0^{1-y_1} \\ &= 3\left(y_1(1 - y_1) + \frac{(1 - y_1)^2}{2} \right) \\ &= \frac{3(1 - y_1^2)}{2}, \quad y_1 \in (0, 1). \end{aligned}$$

□

Example 5.3.4. Let X_1 and X_2 be independent $\mathrm{N}(0,1)$ rvs:

$$f_{\boldsymbol{X}}(x_1,x_2)=\frac{1}{\sqrt{2\pi}}e^{-x_1^2/2}\cdot\frac{1}{\sqrt{2\pi}}e^{-x_2^2/2},\quad x_i\in\mathbb{R}.$$

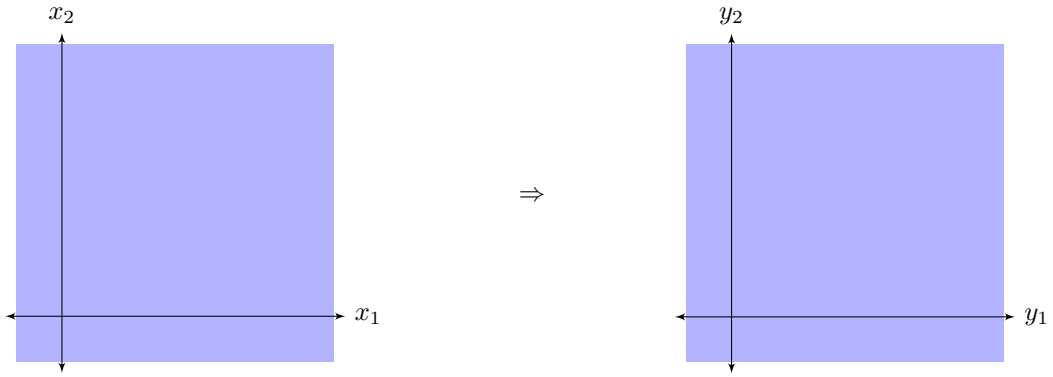
We want to find the distribution of $Y_1=\frac{X_1}{X_2}$, as we did in Example 5.2.3. Now we will get the same result (as expected), but from a different method.

$$\begin{array}{ll}Y_1=X_1/X_2=g_1(X_1,X_2)\\Y_2=X_2=g_2(X_1,X_2)\end{array}\quad\Rightarrow\quad\begin{array}{ll}X_1=Y_1Y_2=h_1(Y_1,Y_2)\\X_2=Y_2=h_2(Y_1,Y_2)\end{array}.$$

The Jacobian is

$$J=\begin{vmatrix}\frac{\partial x_1}{\partial y_1}&\frac{\partial x_1}{\partial y_2}\\\frac{\partial x_2}{\partial y_1}&\frac{\partial x_2}{\partial y_2}\end{vmatrix}=\begin{vmatrix}y_2&y_1\\0&1\end{vmatrix}=y_2,\quad-\infty<y_1<\infty.$$

Our regions of integration are both $\mathbb{R}^2$:



Therefore,

$$f_{\boldsymbol{Y}}(y_1,y_2)=f_{\boldsymbol{X}}(y_1y_2,y_2)\cdot|y_2|=\frac{1}{2\pi}\exp\left(-\frac{(y_1y_2)^2}{2}\right)\exp\left(-\frac{y_2^2}{2}\right)|y_2|,\quad y_1,y_2\in\mathbb{R}$$

and so

$$\begin{aligned}
f_{Y_1}(y_1) &= \int_{S_{Y_2}} f_{Y_1,Y_2}(y_1, y_2)\, dy_2 \\
&= \int_{-\infty}^{\infty} \frac{|y_2|}{2\pi} \exp\left(-\frac{y_2^2(1+y_1^2)}{2}\right)\, dy_2 \\
&= \frac{1}{2\pi}\int_{-\infty}^{0} -y_2 \exp\left(-\frac{y_2^2(1+y_1^2)}{2}\right)\, dy_2 + \frac{1}{2\pi}\int_{0}^{\infty} y_2 \exp\left(-\frac{y_2^2(1+y_1^2)}{2}\right)\, dy_2 \\
&= \frac{1}{\pi}\int_{0}^{\infty} y_2 \exp\left(-\frac{y_2^2(1+y_1^2)}{2}\right)\, dy_2 \\
&= -\frac{1}{\pi(1+y_1^2)}\int_{0}^{\infty} -y_2(1+y_1^2) \exp\left(-\frac{y_2^2(1+y_1^2)}{2}\right)\, dy_2 \\
&= -\frac{1}{\pi(1+y_1^2)}\left[\exp\left(-\frac{y_2^2(1+y_1^2)}{2}\right)\right]_0^{\infty} \\
&= -\frac{1}{\pi(1+y_1^2)}\,(0-1) \\
&= \frac{1}{\pi(1+y_1^2)}, \quad y_1 \in \mathbb{R}.
\end{aligned}$$

which agrees with our previous result. □

Example 5.3.5. Let $X_1, X_2, X_3 \overset{iid}{\sim} \text{Exp}(1)$:

$$f_X(x) = e^{-x}, \quad x > 0.$$

Let $Y_1 = \frac{X_1}{X_1+X_2}$, $Y_2 = \frac{X_1+X_2}{X_1+X_2+X_3}$, and $Y_3 = X_1 + X_2 + X_3$. We want to show that $Y_1 \sim \text{U}(0,1)$, $Y_3 \sim \text{Gamma}(\alpha = 3, \beta = 1)$[2], and that Y_1, Y_2, Y_3 are all independent.

$$\begin{aligned} Y_1 &= \frac{X_1}{X_1+X_2} = g_1(X_1,X_2,X_3) \\ Y_2 &= \frac{X_1+X_2}{X_1+X_2+X_3} = g_2(X_1,X_2,X_3) \\ Y_3 &= X_1+X_2+X_3 = g_3(X_1,X_2,X_3) \end{aligned} \quad \Rightarrow \quad \begin{aligned} X_1 &= Y_1Y_2Y_3 = h_1(Y_1,Y_2,Y_3) \\ X_2 &= -Y_1Y_2Y_3 + Y_2Y_3 = h_2(Y_1,Y_2,Y_3) \\ X_3 &= -Y_2Y_3 + Y_3 = h_3(Y_1,Y_2,Y_3) \end{aligned},$$

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \frac{\partial x_1}{\partial y_3} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \frac{\partial x_2}{\partial y_3} \\ \frac{\partial x_3}{\partial y_1} & \frac{\partial x_3}{\partial y_2} & \frac{\partial x_3}{\partial y_3} \end{vmatrix} = \begin{vmatrix} y_2y_3 & y_1y_3 & y_1y_2 \\ -y_2y_3 & -y_1y_3+y_3 & -y_1y_2+y_2 \\ 0 & -y_3 & -y_2+1 \end{vmatrix} = y_2y_3^2.$$

One can show that $y_1, y_2 \in (0,1)$ and that $y_3 > 0$, and that

$$f_{\boldsymbol{Y}}(y_1,y_2,y_3) = y_2y_3^2e^{-y_3}, \quad y_1,y_2 \in (0,1);\ y_3 > 0.$$

Recall that $y_3 = x_1 + x_2 + x_3$. Therefore,

$$f_{Y_1}(y_1) = \int_0^\infty \int_0^1 y_2y_3^2e^{-y_3}\, dy_2dy_3 = 1, \quad y_1 \in (0,1),$$

$$f_{Y_2}(y_2) = \int_0^\infty \int_0^1 y_2y_3^2e^{-y_3}\, dy_1dy_3 = y_2\int_0^\infty y_3^2e^{-y_3}\, dy_3 = 2y_2, \quad y_2 \in (0,1),$$

$$f_{Y_3}(y_3) = \int_0^1 \int_0^1 y_2y_3^2e^{-y_3}\, dy_1dy_2 = y_3^2e^{-y_3}\int_0^1 y_2\, dy_2 = \frac{1}{2}y_3^2e^{-y_3}, \quad y_3 > 0.$$

Note that $f_{\boldsymbol{Y}}(y_1,y_2,y_3) = f_{Y_1}(y_1)f_{Y_2}(y_2)f_{Y_3}(y_3)$. □

[2] A *Gamma distribution* is one with has the pdf

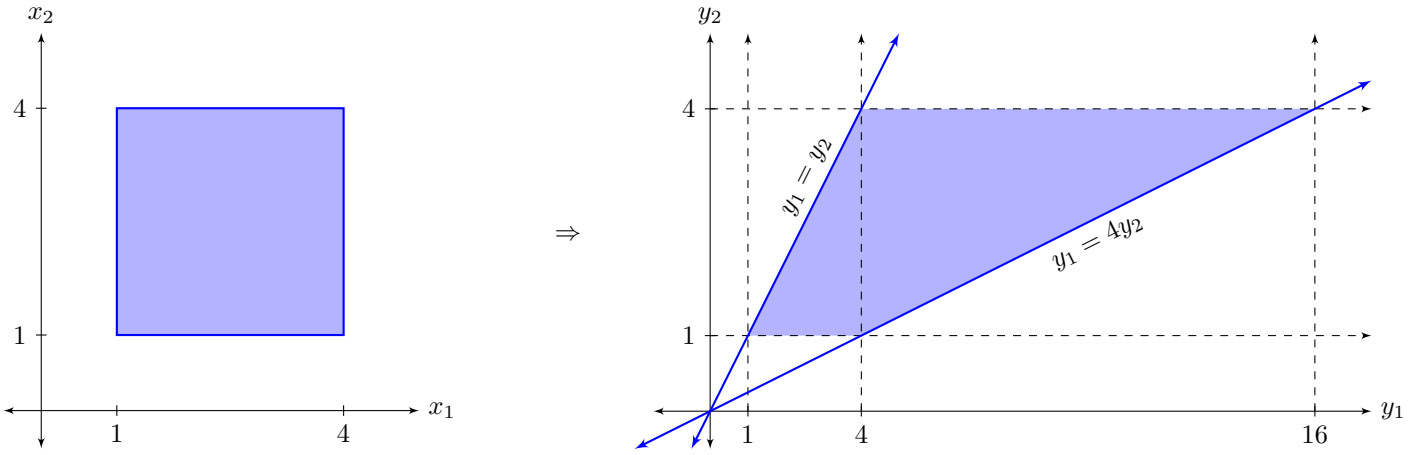
$$f_X(x) = \frac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)}, \quad x > 0.$$

Example 5.3.6. Let $X_1, X_2 \overset{iid}{\sim} \mathrm{U}(1,4)$. Find the pdf of $Y = X_1 X_2$.

Let

$$\begin{array}{ll} Y_1 = X_1X_2 = g_1(X_1,X_2) & \\ Y_2 = X_2 = g_2(X_1,X_2) & \end{array} \quad \Rightarrow \quad \begin{array}{l} X_1 = Y_1/Y_2 = h_1(Y_1,Y_2) \\ X_2 = Y_2 = h_2(Y_1,Y_2) \end{array} .$$

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} \frac{1}{y_2} & -\frac{y_1}{y_2^2} \\ 0 & 1 \end{vmatrix} = \frac{1}{y_2}, \quad y_2 \in (1,4).$$



$$\begin{array}{lllll} 1 \le x_1 \le 4 & \Rightarrow & 1 \le \dfrac{y_1}{y_2} \le 4 & \Rightarrow & y_2 \le y_1 \le 4y_2, \\ 1 \le x_2 \le 4 & \Rightarrow & 1 \le y_2 \le 4. & & \end{array}$$

$$f_{\boldsymbol{X}}(x_1,x_2) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}, \quad 1 \le x_i \le 4,$$

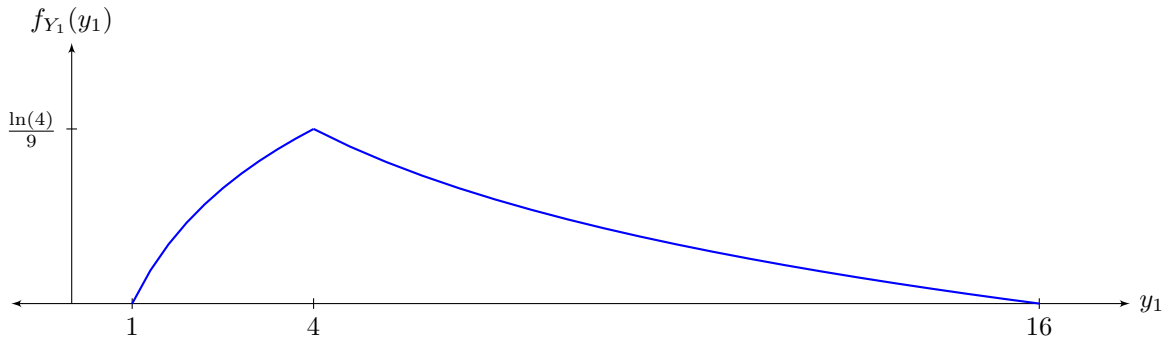
$$f_{\boldsymbol{Y}}(y_1,y_2) = f_{\boldsymbol{X}}(y_1/y_2, y_2) \cdot \frac{1}{y_2} = \frac{1}{9y_2}, \quad y_2 \le y_1 \le 4y_2, \quad y_2 \in [1,4].$$

For $1 \le y_1 \le 4$,

$$f_{Y_1}(y_1) = \int_1^{y_1} \frac{1}{9y_2}\, dy_2 = \frac{1}{9}\ln(y_1),$$

For $4 \le y_1 \le 16$,

$$f_{Y_1}(y_1) = \int_{y_1/4}^{4} \frac{1}{9y_2}\, dy_2 = \frac{1}{9}\left(\ln(4) - \ln\left(\frac{y_1}{4}\right)\right) = \frac{\ln(16) - \ln(y_1)}{9}$$



□

Overall, when calculating the pdf of a function Y of n random variables, there are two approaches to take:

- Use multiple integrals to compute the distribution function $F_Y(y)$, then take its derivative $f_Y(y)$.
- Use Jacobians from n-space to n-space to compute the joint pdf $f_{\boldsymbol{Y}}(y, y_2, \ldots, y_n)$ for some r.v.'s $Y_2, \ldots, Y_n$ that are conveniently defined, then integrate out the $n-1$ r.v.'s

The outcome of each of these two approaches will be the same. We shall illustrate this with one example done via these two different ways.

Example 5.3.7. Let $X_1, X_2 \overset{iid}{\sim} \text{Exp}(\lambda)$. What is the pdf of $Y = X_1 + X_2$?

First note that, since X_1 and X_2 are independent,

$$f_{X_1X_2}(x_1, x_2) = f_X(x_1)f_X(x_2) = \lambda e^{-\lambda x_1}\lambda e^{-\lambda x_2} = \lambda^2 e^{-\lambda(x_1+x_2)}, \qquad x_1, x_2 > 0. \tag{5.3.4}$$

Method 1: Integrals

$$\begin{aligned}
F_Y(y) &= \text{P}(X_1 + X_2 \le y) \\
&= \text{P}(X_2 \le y - X_1) \\
&= \int_0^y \int_0^{y-x_1} \lambda^2 e^{-\lambda x_1 - \lambda x_2}\, dx_2 dx_1 && \text{by a graph of } x_1 \text{ vs. } x_2 \\
&= \int_0^y \lambda e^{-\lambda x_1} \left[-e^{-\lambda x_2}\right]_0^{y-x_1}\, dx_1 \\
&= \int_0^y \lambda e^{-\lambda x_1} \left(1 - e^{-\lambda(y-x_1)}\right)\, dx_1 \\
&= \int_0^y \lambda e^{-\lambda x_1}\, dx_1 - \int_0^y \lambda e^{-\lambda y}\, dx_1 \\
&= \left[-e^{-\lambda x_1}\right]_0^y - \lambda e^{-\lambda y}(y) \\
&= 1 - e^{-\lambda y} - y\lambda e^{-\lambda y}
\end{aligned}$$

Therefore,

$$f_Y(y) = \frac{d}{dy}[F_Y(y)] = \lambda e^{-\lambda y} - y(-\lambda^2 e^{-\lambda y}) - \lambda e^{-\lambda y} = \lambda^2 y e^{-\lambda y}, \quad y > 0,$$

so that by inspection (i.e., by comparison to the formula for the pdf of the Gamma distribution), we can conclude that $Y \sim \text{Gamma}(2, \lambda)$.

Method 2: Jacobians

Let

$$\begin{array}{rcl} Y &=& X_1 + X_2 \\ Z &=& X_2 \end{array} \quad \Leftrightarrow \quad \begin{array}{rcl} X_1 &=& Y - Z \\ X_2 &=& Z \end{array}$$

where we note that $Y > 0$ and $0 < Z < Y$ (since $Z = X_2 > 0$ and $X_1 = Y - Z > 0 \quad \Rightarrow \quad Y > Z$). Thus,

$$J = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1,$$

so that, by 5.3.4,

$$f_{YZ}(y, z) = f_{X_1 X_2}(y - z, z)|J| = \lambda^2 e^{-\lambda(y-z+z)} \cdot 1 = \lambda^2 e^{-\lambda y}, \qquad y > 0, \;\; 0 < z < y.$$

Therefore,

$$f_Y(y) = \int_0^y f_{YZ}(y, z)\, dz = \int_0^y \lambda^2 e^{-\lambda y}\, dz = \lambda^2 y e^{-\lambda y}, \qquad y > 0, \qquad \square$$

which, as expected, is the same answer we reached in Method 1 above.

6 Mathematical Expectation

6.1 General Definitions

Consider a random variable with a probability law $\mathrm{P}_X(\cdot)$. If $g(x)$ is a continuous function of x for $x \in \mathbb{R}$, then we wish to define the *expected value of* $g(X)$ *wrt the probability law* $P_X(\cdot)$, which we will denote as $\mathrm{E}_X[g(X)]$.

Definition 6.1.1. If X is a discrete rv, then

$$\mathrm{E}_X[g(X)] = \sum_{\substack{\text{all } x \ni \\ p_X(x)>0}} g(x)p_X(x).$$

Note that $\mathrm{E}_X[g(X)]$ exists iff

$$\mathrm{E}_X[|g(X)|] = \sum_{\substack{\text{all } x \ni \\ p_X(x)>0}} |g(x)|p_X(x) < \infty,$$

i.e., iff the series defining $\mathrm{E}_X[g(X)]$ is absolutely convergent.

Definition 6.1.2. If X is a continuous rv, then

$$\mathrm{E}_X[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)\ dx.$$

$\mathrm{E}_X[g(X)]$ exists iff

$$\mathrm{E}_X[|g(X)|] = \int_{-\infty}^{\infty} |g(x)|f_X(x)\ dx < \infty.$$

Now consider the case where $g(x) = x$:

Theorem 6.1.1. For any rv X and continuous function $g(\cdot)$ with the rv $Y = g(X)$, $\mathrm{E}_Y[Y] = \mathrm{E}_X[g(X)]$.

Proof: Will not be given at this time. See Freeman Chap. 8. The proof depends on *absolute convergence.* ■

The importance of this theorem is that in order to find the expected value of Y, one need not determine the pdf for Y, since for the continuous case,

$$E_Y[Y] = \int_{-\infty}^{\infty} yf_Y(y)\ dy = \int_{-\infty}^{\infty} g(x)f_X(x)\ dx.$$

Definition 6.1.3. $\mathrm{E}[X]$ is called the *mean* of the rv X (or rather, the probability law associated with X) and is sometimes denoted as μ_X.:

$$\text{If } X \text{ is discrete, then } \mathrm{E}[X] = \mu_X = \sum_{\text{all } x} xp_X(x),$$

$$\text{If } X \text{ is continuous, then } \mathrm{E}[X] = \mu_X = \int_{-\infty}^{\infty} xf_X(x)\ dx.$$

The mean provides a measure of the midpoint of the probability distribution.

Assume that for a rv X that $\mathrm{E}[X]$ exists. If we take a sequence $X_1, \ldots, X_n$ of independent observations of X and form

$$\begin{aligned}
\bar{X}_1 &= X_1, \\
\bar{X}_2 &= \frac{X_1 + X_2}{2}, \\
\bar{X}_3 &= \frac{X_1 + X_2 + X_3}{3}, \\
&\vdots \\
\bar{X}_n &= \frac{X_1 + \cdots + X_n}{n},
\end{aligned}$$

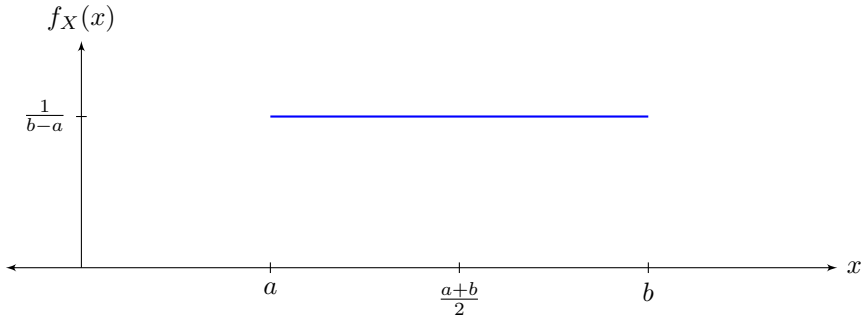
then $\lim_{n\to\infty} \bar{X}_n = \mathrm{E}[X]$ with probability 1.

Example 6.1.1 (Bernoulli).

$$\mathrm{E}[X] = \sum_{x=0}^{1} x p_X(x) = 0 \cdot q + 1 \cdot p = p. \qquad \square$$

Example 6.1.2 (Uniform (a, b)).

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_a^b \frac{x}{b-a}\, dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}.$$



$\square$

Example 6.1.3. Let X be the financial outcome of a game for a player. If

$$\mathrm{E}[X] \begin{cases} = 0 & \text{the game is called fair,} \\ > 0 & \text{the game is called favorable,} \\ < 0 & \text{the game is called unfavorable} \end{cases} .$$

Consider the game of roulette. Use the European wheel with 37 divisions, numbered 0 to 36. Assuming the $\mathrm{P}(\text{The number} = i)) = \frac{1}{37}$ for $i = 0, \ldots, 36$:

$$\begin{cases} 18 \text{ of the numbers are black,} \\ 18 \text{ of the numbers are red,} \\ 1 \text{ is grey.} \end{cases} .$$

A player plays against the house. Two of the systems are

a) Bet on a color (red or black). If the color comes up, the player doubles her stake. If the other color or grey appears, he loses her stake. Let the amount of the bet be \$1.00. What are the expected net financial outcomes of both the player and the house?

If X and Y are the net financial outcomes of the player and the house, respectively, then

$$\mathrm{E}[X] = \frac{18}{37} \cdot (\$1.00) + \frac{19}{37} \cdot (-\$1.00) = \$-\frac{1}{37},$$
$$\mathrm{E}[Y] = \frac{19}{37} \cdot (\$1.00) + \frac{18}{37} \cdot (-\$1.00) = \$+\frac{1}{37}.$$

b) Bet on a number. If that number appears, the player receives 36 times her stake; otherwise, she loses her stake. What are the expected net financial outcomes of both the player and the house?

If X and Y are the net financial outcomes of the player and the house, respectively, then

$$\mathrm{E}[X] = \frac{1}{37} \cdot (\$35.00) + \frac{36}{37} \cdot (-\$1.00) = \$ - \frac{1}{37},$$
$$\mathrm{E}[Y] = \frac{36}{37} \cdot (\$1.00) + \frac{1}{37} \cdot (-\$35.00) = \$ + \frac{1}{37}.$$

Note that in both cases, the expected values for the player and for the house add up to zero (as we would expect).□

Example 6.1.4 (Normal (μ, σ^2)).

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \mathbb{R}.$$

Then

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} \frac{x}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}\, dx \\ &= \int_{-\infty}^{\infty} \frac{\sigma y + \mu}{\sigma\sqrt{2\pi}} e^{-y^2/2}\, \sigma dy && y \equiv \tfrac{x-\mu}{\sigma} \quad \Rightarrow \quad x = \sigma y + \mu,\ dx = \sigma dy \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y e^{-y^2/2}\, dy + \mu \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy}_{=1} && \text{pdf of N}(0,1) \text{ rv} \\ &= \frac{\sigma}{\sqrt{2\pi}} \underbrace{\left[-e^{-y^2/2}\right]_{-\infty}^{\infty}}_{=0} + \mu \cdot 1 \\ &= \mu. \end{aligned}$$

□

Definition 6.1.4. In general, $\mathrm{E}_X[X^n]$ is called the n^{th} *moment* of X. As with $\mathrm{E}[X]$, $\mathrm{E}[X^n]$ exists iff $\mathrm{E}[|X^n|] < \infty$.

Example 6.1.5 (Uniform (a, b)). The second moment of X, where $X \sim \mathrm{U}(a, b)$, is

$$\mathrm{E}[X^2] = \int_a^b x^2 \frac{1}{b-a}\, dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{1}{3(b-a)}(b^3 - a^3) = \frac{b^2+ab+a^2}{3}. \qquad \square$$

The definition of mathematical expectation can be extended to n-dimensional rvs.

Definition 6.1.5. The quantity $\mathrm{E}\left[(X - \mathrm{E}[X])^n\right]$ is the n^{th} *central moment* of X.

Definition 6.1.6. The 2^{nd} central moment of X is called the *variance* of X[1]:

$$\sigma_X^2 = \sigma^2[X] = \mathrm{Var}(X) = \mathrm{E}[(X - \mu_X)^2]$$

Example 6.1.6 (Bernoulli).

$$\mathrm{E}[(X-\mu_X)^2] = \mathrm{E}[(X-p)^2] = \sum_{x=0}^{1}(x-p)^2 p_X(x) = \left(-p^2\right)\cdot q + (1-p)^2 \cdot p = pq(p + (1-p)) = pq. \qquad \square$$

Definition 6.1.7. The positive square root of the variance of X is called the *standard deviation of* X: $\sigma_X = \sqrt{\mathrm{Var}(X)}$.

From Example 6.1.6, the standard deviation of a Bernoulli(p) rv is $\sqrt{pq}$.

The definition of mathematical expectation can be extended to n-dimensional rvs.

Definition 6.1.8. Let $g(\cdot)$ be a function of the n-dimensional rv $\boldsymbol{X} = (X_1, \ldots, X_n)$.. Then

$$\mathrm{E}_{\boldsymbol{X}}[g(\boldsymbol{X})] = \mathrm{E}_{X_1,\ldots,X_n}[g(X_1,\ldots,X_n)] = \begin{cases} \displaystyle\sum_{x_n}\cdots\sum_{x_1} g(x_1,\ldots,x_n)\cdot p_{\boldsymbol{X}}(x_1,\ldots,x_n) \\ \displaystyle\int_{x_n}\cdots\int_{x_1} g(x_1,\ldots,x_n)\cdot f_{\boldsymbol{X}}(x_1,\ldots,x_n)\, dx_1\cdots dx_n \end{cases}.$$

As in the 1-dimensional case, $\mathrm{E}_{\boldsymbol{X}}[g(\boldsymbol{X})]$ exists iff $\mathrm{E}_{\boldsymbol{X}}[|g(\boldsymbol{X})|] < \infty$.

Example 6.1.7. Let the rv $\boldsymbol{X} = (X_1, X_2)$, where $S = \{(x_1, x_2) :\ x_i = 1, \ldots, 6\}$, as in outcomes on a pair of dice. Thus,

$$p_{\boldsymbol{X}}(x_1, x_2) = \frac{1}{36}, \quad (x_1, x_2) \in S.$$

Let $g(x_1, x_2) = x_1 + x_2$. Then

$$\mathrm{E}_{\boldsymbol{X}}[g(\boldsymbol{X})] = \mathrm{E}[X_1 + X_2] = \sum_{x_2=1}^{6}\sum_{x_1=1}^{6}(x_1+x_2)p_{\boldsymbol{X}}(x_1,x_2) = \frac{1}{36}\left[(1+1)+(1+2)+\cdots+(6+6)\right] = 7. \qquad \square$$

[1] In physics, σ^2 is called the *moment of inertia* wrt the line perpendicular to the x-axis and passing through the point $x = \mathrm{E}[X]$, $y = 0$, where $\mathrm{E}[X]$ is the center of gravity of unit mass.

6.2 Properties of Mathematical Expectation

Theorem 6.2.1. $\mathrm{E}_X[c] = c$.

Proof: Without loss of generality, we shall only cover the continuous case; the discrete case is similar.

$$\mathrm{E}_X[c] = \int_{-\infty}^{\infty} c f_X(x)\ dx = c \int_{-\infty}^{\infty} f_X(x)\ dx = c \cdot 1 = c.$$

■

Theorem 6.2.2. $\mathrm{E}_X[cg(X)] = c\mathrm{E}_X[g(X)]$.

Proof: Homework problem! ■

Example 6.2.1 (Normal (μ, σ^2)). Let $g(x) = x$. Then $\mathrm{E}[6g(X)] = 6\mathrm{E}[g(X)] = 6\mathrm{E}[X] = 6\mu$. □

Theorem 6.2.3. $\mathrm{E}[g_1(X) + g_2(X)] = \mathrm{E}[g_1(X)] + \mathrm{E}[g_2(X)]$.

Proof: Without loss of generality, we shall only cover the continuous case; the discrete case is similar.

$$\mathrm{E}_X[g_1(X) + g_2(X)] = \int_{-\infty}^{\infty} (g_1(x) + g_2(x))\, f_X(x)\ dx = \int_{-\infty}^{\infty} g_1(x) f_X(x)\ dx + \int_{-\infty}^{\infty} g_2(x) f_X(x)\ dx = \mathrm{E}_X[g_1(X)] + \mathrm{E}_X[g_2(x)].$$

■

Theorem 6.2.3 can be extended to n functions $\{g_j(x)\}$ by using mathematical induction:

$$\mathrm{E}_X[g_1(X) + \cdots + g_n(X)] = \mathrm{E}_X[g_1(X)] + \cdots + \mathrm{E}_X[g_n(X)].$$

Example 6.2.2. Let $X \sim \mathrm{N}(3, \sigma^2 = 2)$. If $g_1(x) = 5x$ and $g_2(x) = -4$, then

$$\mathrm{E}_X[g_1(X) + g_2(X)] = \mathrm{E}_X[5X] + \mathrm{E}_X[-4] = 5 \cdot 3 + (-4) = 11.$$ □

Theorem 6.2.4. $\mathrm{E}[ag(X) + b] = a\mathrm{E}[g(X)] + b$

Proof: A direct consequence of Theorems 6.2.2 and 6.2.3 above. ■

Note that a special case of Theorem 6.2.4 is $g(X) = X$, which gives us $\mathrm{E}[aX + b] = a\mathrm{E}[X] + b$.

Theorem 6.2.5. If $\mathrm{E}_{X_1}[g_1(X_1)]$ and $\mathrm{E}_{X_2}[g_2(X_2)]$ exist and if $g(x_1, x_2) = g_1(x) + g_2(x)$, then

$$\mathrm{E}_{X_1,X_2}[g(X_1, X_2)] = \mathrm{E}_{X_1,X_2}[g_1(X_1) + g_2(X_2)] = \mathrm{E}_{X_1}[g_1(X_1)] + \mathrm{E}_{X_2}[g_2(X_2)].$$

Proof: Again, we shall only prove the continuous case, since the discrete case is similar.

$$\begin{aligned}
\mathrm{E}_{X_1,X_2}[g(X_1, X_2)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1, x_2) f_{X_1,X_2}(x_1, x_2)\ dx_1 dx_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [g_1(x_1) + g_2(x_2)]\, f_{X_1,X_2}(x_1, x_2)\ dx_1 dx_2 \\
&= \int_{-\infty}^{\infty} g_1(x_1) \left[\int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)\ dx_2\right]\ dx_1 + \int_{-\infty}^{\infty} g_2(x_2) \left[\int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)\ dx_1\right]\ dx_2 \\
&= \int_{-\infty}^{\infty} g_1(x_1) f_{X_1}(x_1)\ dx_1 + \int_{-\infty}^{\infty} g_2(x_2) f_{X_2}(x_2)\ dx_2 \\
&= \mathrm{E}_{X_1}[g_1(X_1)] + \mathrm{E}_{X_2}[g_2(X_2)].
\end{aligned}$$

■

As before, Theorem 6.2.5 can be extended to n functions $\{g_j(x)\}$ by using mathematical induction:

$$\mathrm{E}_{X_1,\ldots,X_n}[g_1(X_1) + \cdots + g_n(X_n)] = \mathrm{E}_{X_1}[g_1(X_1)] + \cdots + \mathrm{E}_{X_n}[g_n(X_n)].$$

We are usually interested in the special case of $g_j(x_j) = x_j$:

$$\mathrm{E}_{X_1,\ldots,X_n}[X_1 + \cdots + X_n] = \mathrm{E}_{X_1}[X_1] + \cdots + \mathrm{E}_{X_n}[X_n].$$

Note the differences and similarities between Theorems 6.2.3 and 6.2.5. Also note that there are no assumptions of independence in either of these theorems.

Example 6.2.3. Suppose $X_1 \sim \mathrm{N}(3, \sigma^2 = 2)$, $X_2 \sim \mathrm{N}(6, \sigma^2 = 1)$ and $X_3 \sim \mathrm{U}(4, 8)$. Then by Theorem 6.2.5,

$$\mathrm{E}_{X_1,X_2,X_3}[X_1 + X_2 + X_3] = \mathrm{E}[X_1] + \mathrm{E}[X_2] + \mathrm{E}[X_3] = 3 + 6 + \frac{4+8}{2} = 15.$$

This is true whether or not X_1, X_2 and X_3 are all independent of each other. □

Theorem 6.2.6. $\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2$

Proof: For simpler notation, let $\mu = \mathrm{E}[X]$. Then

$$\begin{aligned}
\mathrm{E}\left[(X - \mathrm{E}[X])^2\right] &= \mathrm{E}\left[X^2 - 2\mu X + \mu^2\right] \\
&= \mathrm{E}\left[X^2\right] - 2\mu\mathrm{E}[X] + \mu^2 \\
&= \mathrm{E}\left[X^2\right] - 2\mu^2 + \mu^2 \\
&= \mathrm{E}\left[X^2\right] - \mu^2 \\
&= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2.
\end{aligned}$$

■

Example 6.2.4 (Bernoulli). Recall that $\mathrm{E}[X] = p$ and $\mathrm{Var}(X) = pq$. We can check this formula for the variance using Theorem 6.2.6:

$$\mathrm{E}\left[X^2\right] = \sum_{x=0}^{1} x^2 p_X(x) = 0^2 \cdot q + 1^2 \cdot p.$$

Therefore, by Theorem 6.2.6, $\mathrm{Var}(X) = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 = p - p^2 = p(1-p) = pq$. □

Example 6.2.5 (Uniform (a, b)). Recall that $\mathrm{E}[X] = \frac{b+a}{2}$ and $\mathrm{E}\left[X^2\right] = \frac{b^2+ab+a^2}{3}$. Thus,

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 \\
&= \frac{b^2+ab+a^2}{3} - \left(\frac{b+a}{2}\right)^2 \\
&= \frac{4(b^2+ab+a^2) - 3(b^2+2ab+a^2)}{12} \\
&= \frac{b^2-2ab+a^2}{12} \\
&= \frac{(b-a)^2}{12}.
\end{aligned}$$

□

Theorem 6.2.7. $\mathrm{Var}(ag(X)+b) = a^2\mathrm{Var}(g(X))$

Proof: Letting $\mu = g(X)$,

$$\begin{aligned}
\mathrm{Var}(ag(X)+b) &= \mathrm{E}\left[(ag(X)+b-\mathrm{E}[ag(X)+b])^2\right] \\
&= \mathrm{E}\left[(ag(X)+b-a\mu-b)^2\right] && \text{Theorem 6.2.4} \\
&= \mathrm{E}\left[a^2\left(g(X)-\mu\right)^2\right] \\
&= a^2\mathrm{E}\left[(g(X)-\mu)^2\right] && \text{Theorem 6.2.2} \\
&= a^2\mathrm{Var}(g(X)) && \text{Definition 6.1.6}
\end{aligned}$$

■

Note that a special case of Theorem 6.2.7 is $g(X) = X$, which gives us $\mathrm{Var}[aX+b] = a^2\mathrm{Var}[X]$.

For any rv, the existence of σ_X^2 implies the existence of $\mathrm{E}[X]$. In fact, it can be proven that the existence of $\mathrm{E}\left[X^2\right]$ $\Rightarrow \mathrm{E}[X]$ exists. However, the reverse is not true. That is,

$$\mathrm{E}[X] < \infty \quad \not\Rightarrow \quad \mathrm{E}\left[X^2\right] < \infty.$$

Example 6.2.6. This example is from Birnbaum (1962). Let X be a discrete rv with the possible values

$$x_k = k^{-1/2}2^{k/2}, \quad k \in \mathbb{Z}^+$$

with probabilities

$$p_X(x_k) = \frac{1}{2^k}, \quad k \in \mathbb{Z}^+.$$

For the expectation, we have

$$\mathrm{E}[X] = \sum_{k=1}^{\infty} k^{-1/2} 2^{-k/2} < \sum_{k=1}^{\infty} 2^{-k/2} = \sum_{k=1}^{\infty} \left(\frac{1}{\sqrt{2}}\right)^k < \infty$$

since the last series is a convergent geometric progression (since $\frac{1}{\sqrt{2}} < 1$). However, for the variance,

$$\mathrm{Var}(X) = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 = \sum_{k=1}^{\infty} k^{-1} \cdot 2^k \cdot \frac{1}{2^k} - (\mathrm{E}[X])^2 = \sum_{k=1}^{\infty} \frac{1}{k} - (\mathrm{E}[X])^2 = \infty - (\mathrm{E}[X])^2 = \infty,$$

which is a divergent harmonic series less a finite number, and hence an infinite number. □

Theorem 6.2.8. Let $g_1(\cdot)$ and $g_2(\cdot)$ be continuous functions defined over the real line. Then if the rvs X_1 and X_2 are independent,

$$\mathrm{E}_{X_1,X_2}[g_1(X_1) \cdot g_2(X_2)] = \mathrm{E}_{X_1}[g_1(X_1)] \cdot \mathrm{E}_{X_2}[g_2(X_2)].$$

Proof: Continuous case:

$$\begin{aligned}
\mathrm{E}[g_1(X_1)g_2(X_2)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_1(x_1)g_2(x_2) f_{X_1,X_2}(x_1,x_2)\, dx_1 dx_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_1(x_1)g_2(x_2) f_{X_1}(x_1) f_{X_2}(x_2)\, dx_1 dx_2 \qquad \text{independence of } X_1 \text{ and } X_2 \\
&= \int_{-\infty}^{\infty} g_1(x_1) f_{X_1}(x_1)\, dx_1 \int_{-\infty}^{\infty} g_2(x_2) f_{X_2}(x_2)\, dx_2 \\
&= \mathrm{E}_{X_1}[g_1(X_1)] \cdot \mathrm{E}_{X_2}[g_2(X_2)].
\end{aligned}$$

■

Example 6.2.7. Let $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ for $i = 1, 2$ and assume that X_1 and X_2 are independent. By Theorem 6.2.8,

$$\mathrm{E}_{X_1,X_2}\left[6X_1X_2^2\right] = 6\mathrm{E}_{X_1}[X_1]\mathrm{E}_{X_2}\left[X_2^2\right] = 6\mu_1\left(\sigma_2^2 - \mu_2^2\right).$$ □

Example 6.2.8. Let $X_1 \sim \mathrm{N}(3, \sigma^2 = 16)$ be independent of $X_2 \sim \mathrm{U}(4, 8)$. By Theorem 6.2.8,

$$\begin{aligned}
\mathrm{E}_{X_1,X_2}[X_1X_2] &= 3(6) = 18, \\
\mathrm{E}_{X_1,X_2}\left[X_1^2X_2\right] &= \mathrm{E}_{X_1}\left[X_1^2\right]\mathrm{E}_{X_2}\left[X_2\right] = (16-9)(6) = 42.
\end{aligned}$$ □

Theorem 6.2.9. Let (X_1, X_2) be a jointly distributed rv. Then

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{E}\left[(X_1 - \mu_1)(X_2 - \mu_2)\right],$$

where $\mu_i = \mathrm{E}[X_i]$.

Proof:

$$\begin{aligned}
\mathrm{Var}(X_1 + X_2) &= \mathrm{E}\left[(X_1 + X_2 - \mu_1 - \mu_2)^2\right] \\
&= \mathrm{E}\left[(X_1 - \mu_1)^2 + 2(X_1 - \mu_1)(X_2 - \mu_2) + (X_2 - \mu_2)^2\right] \\
&= \mathrm{Var}(X_1) + 2\mathrm{E}\left[(X_1 - \mu_1)(X_2 - \mu_2)\right] + \mathrm{Var}(X_2).
\end{aligned}$$

■

Theorem 6.2.10. If the rvs X_1 and X_2 are independent, then $\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$.

Proof:

$$\mathrm{E}_{X_1,X_2}[(X_1 - \mu_1)(X_2 - \mu_2)] = \mathrm{E}_{X_1}[(X_1 - \mu_1)]\mathrm{E}_{X_2}[(X_2 - \mu_2)] = (\mu_1 - \mu_1)(\mu_2 - \mu_2) = 0.$$

Thus, the proof follows from Theorem 6.2.9. ■

Note that $\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) \quad \not\Rightarrow \quad$ independence of X_1 and X_2.

Example 6.2.9. Let $X_1 = \sin(2\pi U)$ and $X_2 = \cos(2\pi U)$, where $U \sim \mathrm{U}(0, 1)$. Then

$$\begin{aligned}
\mathrm{Var}(X_1) &= \int_0^1 \sin^2(2\pi u)\ du - \left(\int_0^1 \sin(2\pi u)\ du\right)^2 = \frac{1}{2}, \\
\mathrm{Var}(X_2) &= \int_0^1 \cos^2(2\pi u)\ du - \left(\int_0^1 \cos(2\pi u)\ du\right)^2 = \frac{1}{2}, \\
\mathrm{Var}(X_1 + X_2) &= \mathrm{Var}\left(\sin(2\pi U) + \cos(2\pi U)\right) = 1.
\end{aligned}$$

However, X_1 and X_2 are not independent, since $f_{X_1,X_2}(x_1, x_2) \neq f_{X_1}(x_1) f_{X_2}(x_2)$. □

6.2.1 Miscellaneous Definitions

Definition 6.2.1. The *covariance* of X_1 and X_2 is

$$\mathrm{E}_{X_1,X_2}[(X_1 - \mu_1)(X_2 - \mu_2)].$$

Again, covariance $= 0 \not\Rightarrow$ independence.

Definition 6.2.2. The *median* of an rv X is a number Me_X such that

$$\mathrm{P}(X < Me_X) \leq \frac{1}{2}, \quad \mathrm{P}(X > Me_X) \leq \frac{1}{2}.$$

Note that in general, median $\neq$ mean.

Example 6.2.10. Let $X \sim \mathrm{Exp}(1)$:

$$f_X(x) = e^{-x}, \quad x > 0.$$

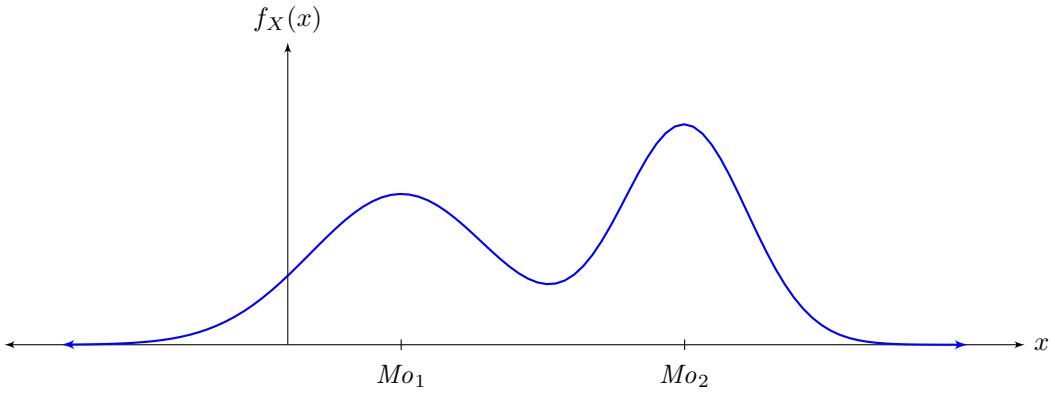
As shown previously, the mean $= \mathrm{E}[X] = 1$. The median of X, Me_X, can be computed as follows:

$$\frac{1}{2} = \int_0^{Me_X} e^{-x}\ dx = \left[-e^{-x}\right]_0^{Me_X} = 1 - e^{-Me_X} \quad \Rightarrow \quad e^{-Me_X} = \frac{1}{2} \quad \Rightarrow \quad Me_X = \ln(2) \approx 0.693. \qquad \square$$

Definition 6.2.3. The *mode* of the rv X is a number Mo_X such that the pdf/pmf has a relative maximum at Mo_X.

In general, a pdf/pmf is called *unimodal* if it has just one mode and *bimodal* if it has two modes.

Example 6.2.11. This is an example of a bimodal distribution:



□

Definition 6.2.4. If σ_X^2 exists, then $Z = \frac{X - \mathrm{E}[X]}{\sigma_X}$ is the *standardized* or *normalized* rv corresponding to X:

$$\mathrm{E}[Z] = 0, \quad \mathrm{Var}(Z) = 1.$$

6.3 Moment Generating Functions and Characteristic Functions

6.3.1 Moment Generating Functions

Definition 6.3.1. If there exists a positive number $h > 0$ such that for $t \in (-h, h)$, $\mathrm{E}_X\left[e^{tX}\right]$ exists (i.e., the integral or sum is absolutely convergent for $t \in (-h, h)$), then

$$M_X(t) = \mathrm{E}_X\left[e^{tX}\right]$$

is called the *moment generating function* of the rv X, which is sometimes abbreviated as "mgf".

Example 6.3.1. Let $X \sim \mathrm{Exp}(\lambda)$:

$$f_X(x) = \lambda e^{-\lambda x}, \quad x > 0.$$

Then

$$\begin{aligned}
M_X(t) &= \mathrm{E}_X\left[e^{tX}\right] \\
&= \int_0^\infty e^{tx} \lambda e^{-\lambda x}\, dx \\
&= \int_0^\infty \lambda e^{-x(\lambda - t)}\, dx \\
&= \frac{\lambda}{\lambda - t} \cdot \underbrace{\int_0^\infty (\lambda - t) e^{-x(\lambda - t)}\, dx}_{=1} && \text{pdf of } \mathrm{Exp}(\lambda - t) \text{ distribution} \\
&= \frac{\lambda}{\lambda - t}, \quad t < \lambda.
\end{aligned}$$

□

Not every distribution has a mgf, as we shall see in Example 6.3.2. When it exists, however, the mgf is unique and completely determines the distribution of the rv. This is a very useful property, as well be demonstrated shortly.

The uniqueness of the mgf is based on the theory of transforms (LaPlace Transforms). Indeed, there is a 1-1 correspondence between the following functions:

$$F_X(\cdot) \quad \Leftrightarrow \quad \left\{ \begin{array}{c} f_X(\cdot) \\ p_X(\cdot) \end{array} \right\} \quad \Leftrightarrow \quad \mathrm{P}_X(\cdot) \quad \Leftrightarrow \quad M_X(\cdot).$$

If, for example, $M_X(t) = \frac{1}{(1-t)^2},\ t < 1$, then by definition of the mgf,

$$\frac{1}{(1-t)^2} = \int_{-\infty}^{\infty} e^{tx} f_X(x)\ dx, \quad t < 1.$$

From the theory of LaPlace Transforms (and the fact that a transform for a pdf is unique), we find that

$$f_X(x) = xe^{-x}, \quad x > 0.$$

Example 6.3.2. Let[2]

$$p_X(x) = \frac{6}{\pi^2 x^2}, \quad x \in \mathbb{Z}^+.$$

In the mgf exists, then

$$M_X(t) = \mathrm{E}\left[e^{tX}\right] = \sum_{x=1}^{\infty} e^{tx} \frac{6}{\pi^2 x^2}.$$

The ratio test can be used to show that this series diverges of $t > 0$. Thus, $\nexists h > 0 \ \ni\ \forall t \in (-h, h)$, $M_X(t)$ exists. Accordingly, this pmf $p_X(\cdot)$ does not have a mgf. □

Where does the name *moment generating function* come from?

The existence of $M_X(t)$ for $t \in (-h, h) \quad \Rightarrow \quad$ derivatives of all orders exist at $t = 0$:

$$\frac{dM_X(t)}{dt} \equiv M_X'(t) = \begin{cases} \displaystyle\int_{-\infty}^{\infty} xe^{tx} f_X(x)\ dx & X \text{ is continuous} \\ \displaystyle\sum_x xe^{tx} p_X(x) & X \text{ is discrete} \end{cases}.$$

Let $t = 0$. Then

$$M_X'(0) = \left\{ \begin{array}{ll} \displaystyle\int_{-\infty}^{\infty} x f_X(x)\ dx & X \text{ is continuous} \\ \displaystyle\sum_x x p_X(x) & X \text{ is discrete} \end{array} \right\} = \mathrm{E}_X[X].$$

Similarly,

$$\frac{d^2 M_X(t)}{dt^2} \equiv M_X''(t) = \begin{cases} \displaystyle\int_{-\infty}^{\infty} x^2 e^{tx} f_X(x)\ dx & X \text{ is continuous} \\ \displaystyle\sum_x x^2 e^{tx} p_X(x) & X \text{ is discrete} \end{cases}.$$

[2]Recall that $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$.

so that

$$M_X''(0) = \left\{ \begin{array}{ll} \displaystyle\int_{-\infty}^{\infty} x^2 f_X(x)\, dx & X \text{ is continuous} \\ \displaystyle\sum_x x^2 p_X(x) & X \text{ is discrete} \end{array} \right\} = \mathrm{E}_X[X^2].$$

Thus, $\mathrm{Var}(X) = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 = M_X''(0) - (M_X'(0))^2$. Generally, $M_X^{(m)}(0) = \mathrm{E}\left[X^m\right]$, and the derivatives of $M_X(t)$ generate the moments of the distribution of X.

Example 6.3.3. Let $X \sim \mathrm{Exp}(\lambda)$. Then for $t < \lambda$,

$$\begin{aligned} M_X(t) &= \frac{\lambda}{\lambda - t} = \left(1 - \frac{t}{\lambda}\right)^{-1} \\ M_X'(t) &= -\left(1 - \frac{t}{\lambda}\right)^{-2}\left(-\frac{1}{\lambda}\right) && \Rightarrow \quad M_X'(0) = \frac{1}{\lambda} = \mathrm{E}\left[X\right] \\ M_X''(t) &= 2\left(1 - \frac{t}{\lambda}\right)^{-3}\left(-\frac{1}{\lambda}\right)^2 = \frac{2}{\lambda^2}\left(1 - \frac{t}{\lambda}\right)^{-3} && \Rightarrow \quad M_X''(0) = \frac{2}{\lambda^2} = \mathrm{E}\left[X^2\right] \end{aligned}$$

so that $\mathrm{Var}(X) = M_X''(0) - (M_X'(0))^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$. For the general term of the mgf,

$$M_X^{(m)}(t) = \frac{m!}{\lambda^m}\left(1 - \frac{t}{\lambda}\right)^{-(m+1)} \qquad \Rightarrow \quad M_X^{(m)}(0) = \frac{m!}{\lambda^m} = \mathrm{E}\left[X^m\right]. \qquad \square$$

Theorem 6.3.1. Let X_1 and X_2 be independent rvs whose mgfs exist. Then for $Y = X_1 + X_2$,

$$M_Y(t) = M_{X_1}(t) \cdot M_{X_2}(t).$$

Proof:

$$M_Y(t) = \mathrm{E}_Y\left[e^{tY}\right] = \mathrm{E}_{X_1X_2}\left[e^{t(X_1+X_2)}\right] \underset{*}{=} \mathrm{E}_{X_1X_2}\left[e^{tX_1}e^{tX_2}\right] = \mathrm{E}_{X_1}\left[e^{tX_1}\right]\mathrm{E}_{X_2}\left[e^{tX_2}\right] = M_{X_1}(t) \cdot M_{X_2}(t).$$

where we use Theorem 6.2.8 in step $*$. ■

By induction, we can prove (not shown here) that if $X_1, \ldots, X_n$ are independent rvs and if $Y = Y_1 + \cdots + Y_n$, then

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t).$$

Example 6.3.4. Let $X_i \overset{ind}{\sim} \mathrm{P}(\lambda_i)$, $i = 1, 2$:

$$p_{X_i}(x_i) = \frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}, \quad x_i \in \mathbb{Z}^*.$$

Then

$$M_{X_i}(t) = \sum_{x=0}^{\infty} \frac{e^{tx}\lambda_i^x e^{-\lambda_i}}{x!} = e^{-\lambda_i}\sum_{x=0}^{\infty}\frac{(e^t\lambda_i)^x}{x!} = e^{-\lambda_i}e^{\lambda_i e^t} = e^{\lambda_i(e^t-1)}.$$

Let $Y = X_1 + X_2$. Then

$$M_Y(t) = M_{X_1}(t)M_{X_2}(t) = e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1))}.$$

This is the mgf for a Poisson rv with parameter $\lambda_1 + \lambda_2$. By the uniqueness of the mgf, Y must then have a Poisson distribution with parameter $\lambda_1 + \lambda_2$. $\square$

Theorem 6.3.2. If $Y = aX + b$, where $a, b \in \mathbb{R}$ and if the mgf's exist, then

$$M_Y(t) = e^{bt} M_X(at).$$

Proof:

$$M_Y(t) = \mathrm{E}_Y\left[e^{tY}\right] = \mathrm{E}_X\left[e^{t(aX+b)}\right] = \mathrm{E}_X\left[e^{atX+bt}\right] = \mathrm{E}_X\left[e^{bt}e^{atX}\right] = e^{bt}\mathrm{E}_X\left[e^{atX}\right] = e^{bt}M_X(at).$$

■

Example 6.3.5. Let $X_i \overset{ind}{\sim} \mathrm{N}(\mu_i, \sigma_i^2)$, $i = 1, 2$. It can be shown (not here) that $M_{X_i}(t) = \exp(\mu_i t + \sigma_i^2 t^2/2)$.

1. Let $Y = X_1 + X_2$. Then
$$M_Y(t) = M_{X_1}(t) \cdot M_{X_2}(t) = \exp(\mu_1 t + \sigma_1^2 t^2/2) \cdot \exp(\mu_2 t + \sigma_2^2 t^2/2) = \exp((\mu_1 + \mu_2)t + (\sigma_1^2 + \sigma_2^2)t^2/2).$$
This is the mgf for a $\mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ rv.
2. Let $Y = a_1X_1 + a_2X_2$. Then
$$\begin{aligned} M_Y(t) &= M_{a_1X_1}(t) \cdot M_{a_2X_2}(t) \\ &= M_{X_1}(a_1t) \cdot M_{X_2}(a_2t) \\ &= \exp(\mu_1 a_1 t + \sigma_1^2 a_1^2 t^2/2) \cdot \exp(\mu_2 a_2 t + \sigma_2^2 a_2^2 t^2/2) \\ &= \exp((a_1\mu_1 + a_2\mu_2)t + (a_1^2\sigma_1^2 + a_2^2\sigma_2^2)t^2/2). \end{aligned}$$
This is the mgf for a $\mathrm{N}(a_1\mu_1 + a_2\mu_2, a_1^2\sigma_1^2 + a_2^2\sigma_2^2)$ rv. □

Example 6.3.6. Let $X_1 \sim \mathrm{N}(-2, \sigma^2 = 9)$ and $X_2 \sim \mathrm{N}(1, \sigma^2 = 16)$.

1. Let $Y = X_1 + X_2$. then $Y \sim \mathrm{N}(-2 + 1, \sigma^2 = 9 + 16) = \mathrm{N}(-1, \sigma^2 = 25)$.
2. Let $Y = 2X_1 - 3X_2$. Then $Y \sim \mathrm{N}(2 \cdot (-2) - 3 \cdot 1, \sigma^2 = 2^2 \cdot 9 + 3^2 \cdot 16) = \mathrm{N}(-7, \sigma^2 = 180)$. □

Theorem 6.3.3. Let $X_1, \ldots, X_n \overset{ind}{\sim} \mathrm{N}(\mu_j, \sigma_j^2)$. Let $k_j \in \mathbb{R}$ for $j = 1, \ldots, n$. Then

$$Y = k_1X_1 + \cdots + k_nX_n \sim \mathrm{N}\left(\sum_{j=1}^n k_j\mu_j,\ \sigma^2 = \sum_{j=1}^n k_j^2\sigma_j^2\right).$$

Proof: Homework problem. ■

Example 6.3.7. Let $X_1, \ldots, X_{10} \overset{iid}{\sim} \mathrm{N}(2, 9)$. What is the distribution of $Y = \bar{X}_{10} = \frac{X_1 + \cdots + X_{10}}{10}$?

Here $k_j = \frac{1}{10}$ $\forall j$, and

$$\begin{aligned} \sum_{j=1}^{10} k_j\mu_j &= \sum_{j=1}^{10} \frac{1}{10} \cdot 2 = 10 \cdot \frac{1}{10} \cdot 2 = 2, \\ \sum_{j=1}^{10} k_j^2\sigma_j^2 &= \sum_{j=1}^{10} \left(\frac{1}{10}\right)^2 \cdot 9 = 10 \cdot \left(\frac{1}{10}\right)^2 \cdot 9 = \frac{9}{10} = 0.9, \end{aligned}$$

so that $Y = \bar{X}_{10} \sim \mathrm{N}(2, \sigma^2 = 0.9)$. □

6.3.2 Characteristic Functions

Definition 6.3.2. If X is an rv, then $\mathrm{E}_X\left[e^{itX}\right] = \phi_X(t)$ is called the *characteristic function* for the rv X.

The main advantage of the characteristic function over the moment generating function is that the characteristic function *always* exists. First we state a few facts about complex numbers:

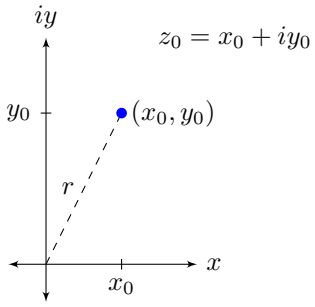
A complex number z consists of a real component x and an imaginary component y:

$$z = x + iy$$

The *complex plane* $\mathbb{C}$ consists of the xy plane representing the real and imaginary components.

Facts:

1. $|z| = |x + iy| = \sqrt{x^2 + y^2}$.
2. $x = r\cos(\theta)$, $y = r\sin(\theta)$.
3. $z = r\cos(\theta) + ir\sin(\theta) = r\left(\cos(\theta) + i\sin(\theta)\right)$.
4. $|z| = r\sqrt{\cos^2(\theta) + \sin^2(\theta)} = r$.
5. For $e^{i\theta} = \cos(\theta) + i\sin(\theta)$, $e^{i\theta} = 1$ for $\theta = 0$.
6. DeMoivre's Theorem: $\left(e^{i\theta}\right)^n = e^{in\theta} = \cos(n\theta) + i\sin(n\theta)$.



Theorem 6.3.4. $\phi_X(t)$ always exists $\forall t$.

Proof:

$$\forall x \in \mathbb{R},\ \left|e^{itx}\right| = \left|\cos(tx) + i\sin(tx)\right| = \sqrt{\cos^2(tx) + \sin^2(tx)} = \sqrt{1} = 1 \quad \Rightarrow \quad \mathrm{E}\left[\left|e^{itX}\right|\right] = \mathrm{E}[1] = 1 < \infty.$$

■

Example 6.3.8. Let $X \sim \mathrm{Exp}(\lambda)$. Then

$$\phi_X(t) = \mathrm{E}_X\left[e^{itX}\right] = \int_0^\infty \lambda e^{itx} e^{-\lambda x}\, dx = \frac{\lambda}{\lambda - it} = \left(1 - \frac{it}{\lambda}\right)^{-1}.$$

Recall that $M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-1}$. □

Example 6.3.9. Let $X \sim \mathrm{P}(\lambda)$. Then

$$\phi_X(t) = \sum_{x=0}^{\infty} e^{itx} \frac{\lambda^x e^{-\lambda}}{x!} = e^{\lambda\left(e^{it} - 1\right)},$$
$$M_X(t) = e^{\lambda\left(e^{t} - 1\right)}.$$ □

From their definitions, we might assume that $\phi_X(t) = M_X(it)$ when $M_X(\cdot)$ exists. This is in fact true, but the proof is not simple, as it depends on analytic concentration from complex variable theory. Thus, we state it here without proof.

As for mfgs, we state the following theorems without proof:

Theorem 6.3.5. If two rvs X and Y have the same characteristic functions, then $F_X(x) = F_Y(x)\ \forall x \in \mathbb{R}$.

$\exists$ a 1-1 correspondence among

$$\phi_X(\cdot) \quad \Leftrightarrow \quad F_X(\cdot) \quad \Leftrightarrow \quad \left\{ \begin{array}{c} f_X(\cdot) \\ p_X(\cdot) \end{array} \right\} \quad \Leftrightarrow \quad \mathrm{P}_X(\cdot) \quad \Leftrightarrow \quad M_X(\cdot) \quad \text{(if it exists)}.$$

Theorem 6.3.6. Let $X_1, \ldots, X_n$ be independent real-valued rvs. If $Y = X_1 + \cdots + X_n$, then $\phi_X(t) = \phi_{X_1}(t) \cdots \phi_{X_n}(t)$.

Theorem 6.3.7. If X is a real-valued rv and $a, b \in \mathbb{R}$, then $\phi_{aX+b}(t) = e^{itb}\phi_X(at)$.

Theorem 6.3.8. If the rv X has an n^{th} moment, then $\mathrm{E}[X^n] = \dfrac{1}{i^n}\phi_X^{(n)}(0)$.

7 Jointly Distributed Random Variables, Continued

7.1 Conditional Distribution

First we will review some basic facts about the jointly distributed rvs. Let $f_{\boldsymbol{X}}(x_1, x_2)$ be the joint pdf for rvs X_1 and X_2. Then

- $\mathrm{P}(a \le X_1 \le b,\ -\infty < X_2 \le c) = \displaystyle\int_{-\infty}^{c}\int_{a}^{b} f_{\boldsymbol{X}}(x_1, x_2)\ dx_1 dx_2.$
- $F_{\boldsymbol{X}}(x_1, x_2) = \displaystyle\int_{-\infty}^{x_2}\int_{-\infty}^{x_1} f_{\boldsymbol{X}}(y_1, y_2)\ dy_1 dy_2.$
- $f_{X_1}(x_1) = \displaystyle\int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, x_2)\ dx_2, \quad f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, x_2)\ dx_1.$

Likewise, if $p_{\boldsymbol{X}}(x_1, x_2)$ is the joint pmf for the rvs X_1 and X_2, then

- $\mathrm{P}(a < X_1 \le b,\ c \le X_2 \le d) = \displaystyle\sum_{x_2=c}^{d}\sum_{x_1=a+1}^{b} p_{\boldsymbol{X}}(x_1, x_2).$
- $F_{\boldsymbol{X}}(x_1, x_2) = \displaystyle\sum_{y_2=-\infty}^{x_2}\sum_{y_1=-\infty}^{x_1} p_{\boldsymbol{X}}(y_1, y_2).$
- $p_{X_1}(x_1) = \displaystyle\sum_{x_2=-\infty}^{\infty} p_{\boldsymbol{X}}(x_1, x_2), \quad p_{X_2}(x_2) = \sum_{x_1=-\infty}^{\infty} p_{\boldsymbol{X}}(x_1, x_2).$

We now want to discuss the notion of a conditional pdf. We will consider two cases: The discrete and the continuous.

7.1.1 Discrete Case

Let X_1 and X_2 be jointly distributed rvs with pmf $p_{\boldsymbol{X}}(x_1, x_2)$ and marginal pdfs $p_{X_1}(x_1)$ and $p_{X_2}(x_2)$. Let

$$\begin{aligned} A_1 &= \{(x_1, x_2) :\ x_1 = a_1,\ -\infty < x_2 < \infty\} \quad \Rightarrow \quad \mathrm{P}(A_1) = \mathrm{P}(X_1 = a_1) = p_{X_1}(a_1), \\ A_2 &= \{(x_1, x_2) :\ x_2 = a_2,\ -\infty < x_1 < \infty\} \quad \Rightarrow \quad \mathrm{P}(A_2) = \mathrm{P}(X_2 = a_2) = p_{X_2}(a_2). \end{aligned}$$

If we know that $p_{X_1}(a_1) > 0$, then we know that

$$\mathrm{P}(X_2 = a_2 \mid X_1 = a_1) = \mathrm{P}(A_2|A_1) = \frac{\mathrm{P}(A_2 A_1)}{\mathrm{P}(A_1)} = \frac{\mathrm{P}(X_1 = a_1,\ X_2 = a_2)}{\mathrm{P}(X_1 = a_1)} = \frac{p_{\boldsymbol{X}}(a_1, a_2)}{p_{X_1}(a_1)}.$$

Note that the quantity $\frac{p_{\boldsymbol{X}}(x_1, x_2)}{p_{X_1}(x_1)}$, for x_1 held constant, is a pmf, since

$$\frac{p_{\boldsymbol{X}}(x_1, x_2)}{p_{X_1}(x_1)} \ge 0 \quad \forall x_2 \in \mathbb{R},$$

and

$$\sum_{x_2=-\infty}^{\infty} \frac{p_{\boldsymbol{X}}(x_1, x_2)}{p_{X_1}(x_1)} = \frac{1}{p_{X_1}(x_1)} \sum_{x_2=-\infty}^{\infty} p_{\boldsymbol{X}}(x_1, x_2) = \frac{1}{p_{X_1}(x_1)} p_{X_1}(x_1) = 1.$$

The *conditional pmf* of X_2 given $X_1 = x_1$ is denoted as $p_{X_2|X_1}(x_2|x_1)$ and defined as the quantity

$$p_{X_2|X_1}(x_2|x_1) = \frac{p_{\boldsymbol{X}}(x_1, x_2)}{p_{X_1(x_1)}} \quad \text{for } p_{X_1}(x_1) > 0.$$

Similarly, $p_{X_1|X_2}(x_1|x_2) = \frac{p_{\boldsymbol{X}}(x_1,x_2)}{p_{X_2(x_2)}}$ for $p_{X_2}(x_2) > 0$. Note that, as with any pmf,

$$\mathrm{P}(a \le X_1 \le b \mid X_2 = x_2) = \sum_{x_1=a}^{b} p_{X_1|X_2}(x_1|x_2),$$
$$\mathrm{P}(c \le X_2 \le d \mid X_1 = x_1) = \sum_{x_2=c}^{d} p_{X_2|X_1}(x_2|x_1).$$

Example 7.1.1. Let

$$p_{\boldsymbol{X}}(x_1, x_2) = \frac{x_1 + x_2}{21}, \quad x_1 \in \{1, 2, 3\},\ x_2 \in \{1, 2\}.$$

Then

$$p_{X_1}(x_1) = \sum_{x_2=1}^{2} \frac{x_1 + x_2}{21} = \frac{1}{21}\left[(x_1 + 1) + (x_1 + 2)\right] = \frac{1}{21}\left(2x_1 + 3\right),$$
$$p_{X_2}(x_2) = \sum_{x_1=1}^{3} \frac{x_1 + x_2}{21} = \frac{1}{21}\left[(1 + x_2) + (2 + x_2) + (3 + x_2)\right] = \frac{1}{21}\left(3x_1 + 6\right).$$

Are X_1 and X_2 independent?

Recall that we may compute $\mathrm{P}(X_1 = 2)$ from the joint or marginal pmfs. That is,

$$\mathrm{P}(X_1 = 2) = \sum_{x_2=1}^{2} p_{\boldsymbol{X}}(2, x_2) = p_{\boldsymbol{X}}(2, 1) + p_{\boldsymbol{X}}(2, 2) = \frac{3}{21} + \frac{4}{21} = \frac{7}{21},$$
$$\mathrm{P}(X_1 = 2) = p_{X_1}(2) = \frac{2(2) + 3}{21} = \frac{7}{21}.$$

The conditional pmfs are given by

$$p_{X_1|X_2}(x_1|x_2) = \frac{p_{\boldsymbol{X}}(x_1, x_2)}{p_{X_2}(x_2)} = \frac{(x_1 + x_2)/21}{(3x_2 + 6)/21} = \frac{x_1 + x_2}{3x_2 + 6}, \quad x_1 \in \{1, 2, 3\};$$
$$p_{X_2|X_1}(x_2|x_1) = \frac{p_{\boldsymbol{X}}(x_1, x_2)}{p_{X_1}(x_1)} = \frac{(x_1 + x_2)/21}{(2x_1 + 3)/21} = \frac{x_1 + x_2}{2x_1 + 3}, \quad x_2 \in \{1, 2\}.$$ □

7.1.2 Continuous Case

If X_1 and X_2 are continuous rvs with joint pdf $f_{\boldsymbol{X}}(x_1, x_2)$ and marginal pdfs $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$, then for $f_{X_1}(x_1) > 0$, we define

$$f_{X_2|X_1}(x_2|x_1) = \frac{f_{\boldsymbol{X}}(x_1, x_2)}{f_{X_1}(x_1)}.$$

This definition is motivated by (7.1.1) for discrete rvs. The fact that $f_{X_2|X_1}(x_2|x_1)$ is a pdf is easily shown, since for $f_{X_1}(x_1) > 0$,

$$f_{X_2|X_1}(x_2|x_1) = \frac{f_{\boldsymbol{X}}(x_1, x_2)}{f_{X_1}(x_1)} \ge 0 \quad \forall x_2 \in \mathbb{R},$$

and

$$\int_{-\infty}^{\infty} f_{X_2|X_1}(x_2|x_1)\ dx_2 = \int_{-\infty}^{\infty} \frac{f_{\boldsymbol{X}}(x_1,x_2)}{f_{X_1}(x_1)}\ dx_2 = \frac{1}{f_{X_1}(x_1)} \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1,x_2)\ dx_2 = \frac{1}{f_{X_1}(x_1)} f_{X_1}(x_1) = 1.$$

Furthermore,

$$\mathrm{P}\left(a \le X_2 \le b \mid X_1 = x_1\right) = \int_a^b f_{X_2|X_1}(x_2|x_1)\ dx_2.$$

Example 7.1.2. Let X_1 and X_2 have the joint pdf

$$f_{\boldsymbol{X}}(x_1,x_2) = 2, \quad 0 < x_1 < x_2 < 1.$$

The marginal pdfs are

$$f_{X_2}(x_2) = \int_0^{x_2} 2\ dx_1 = 2x_2, \quad x_2 \in (0,1);$$
$$f_{X_1}(x_1) = \int_{x_1}^1 2\ dx_2 = 2(1-x_1), \quad x_1 \in (0,1).$$

The conditional pdf of X_1 given $X_2 = x_2$ is

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{\boldsymbol{X}}(x_1,x_2)}{f_{X_2}(x_2)} = \frac{2}{2x_2} = \frac{1}{x_2}, \quad 0 < x_1 < x_2 < 1.$$

Therefore,

$$\mathrm{P}\left(0 < X_1 < \tfrac{1}{2} \mid x_2 = \tfrac{3}{4}\right) = \int_0^{\frac{1}{2}} f_{X_1|X_2}\left(x_1\ \tfrac{3}{4}\right)\ dx_1 = \int_0^{\frac{1}{2}} \frac{1}{3/4}\ dx_1 = \int_0^{\frac{1}{2}} \frac{4}{3}\ dx_1 = \frac{2}{3},$$
$$\mathrm{P}\left(0 < X < \tfrac{1}{2}\right) = \int_0^{\frac{1}{2}} f_{X_1}(x_1)\ dx_1 = \int_0^{\frac{1}{2}} 2(1-x_1)\ dx_1 = \frac{3}{4}.$$ □

7.2 Conditional Expectation

Let $g(X_2)$ be a function of X_2. Then we define

$$\mathrm{E}_{X_2|X_1}\left[g(X_2)|X_1 = x_1\right] = \int_{-\infty}^{\infty} g(x_2) f_{X_2|X_1}(x_2|x_1)\ dx_2$$

as the *conditional expectation* of $g(X_2)$ given $X_1 = x_1$. If they exist, $\mathrm{E}[X_2|X_1 = x_1]$ is the *mean* and $\mathrm{E}\left[\left(X_2 - \mathrm{E}[X_2|X_1 = x_1]\right)^2\right]$ is the *variance* of the conditional distribution of X_2 given $X_1 = x_1$.

Similarly, we define the mean and variance of the conditional distribution of X_1 given $X_2 = x_2$.[1]

[1]Note that $\mathrm{E}[X_1|X_2 = x_2]$ is called the *regression* of X_1 on X_2, and similarly $\mathrm{E}[X_2|X_1 = x_1]$ is the regression of X_2 on X_1.

Example 7.2.1 (Continuation of Example 7.1.2).

$$f_{\boldsymbol{X}}(x_1, x_2) = 2, \quad 0 < x_1 < x_2 < 1;$$

$$f_{X_1}(x_1) = 2(1 - x_1), \quad x_1 \in (0, 1); \qquad f_{X_2}(x_2) = 2x_2, \quad x_2 \in (0, 1).$$

$$f_{X_1|X_2}(x_1|x_2)\ dx_1 = \frac{1}{x_2}, \quad 0 < x_1 < x_2 < 1.$$

$$\mathrm{E}[X_1|x_2] = \int_{-\infty}^{\infty} x_1 f_{X_1|X_2}(x_1|x_2)\ dx_1 = \int_0^{x_2} x_1 \cdot \frac{1}{x_2}\ dx_1 = \frac{x_2}{2}, \quad 0 < x_2 < 1.$$
$$\mathrm{E}\left[(X_1 - \mathrm{E}[X_1|x_2])^2\,|x_2\right] = \int_0^{x_2} \left(x_1 - \tfrac{x_2}{2}\right)^2 \cdot \frac{1}{x_2}\ dx_1 = \frac{x_2^2}{12}, \quad 0 < x_2 < 1. \qquad \square$$

All the definitions given above can be generalized to the case of n variables. We will discuss this generalization for continuous rvs. The generalization for discrete rvs is analogous.

Suppose we have joint pdf $f_{\boldsymbol{X}}(x_1, \ldots, x_n)$.

$$\begin{aligned}
f_{X_1} &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, \ldots, x_n)\ dx_2 dx_3 \cdots dx_n, \\
f_{X_2} &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, \ldots, x_n)\ dx_1 dx_3 \cdots dx_n, \\
&\vdots \\
f_{X_i} &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, \ldots, x_n)\ dx_1 dx_2 \ldots dx_{i-1} dx_{i+1} \cdots dx_n.
\end{aligned}$$

Furthermore,

$$f_{X_1,X_4,X_5}(x_1, x_4, x_5) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\boldsymbol{X}}(x_1, \ldots, x_n)\ dx_2 dx_3 dx_6 dx_7 \cdots dx_n.$$

If $f_{X_1}(x_1) > 0$, then we define

$$f_{X_2,\ldots,X_n|X_1}(x_2, \ldots, x_n|x_1) = \frac{f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)}{f_{X_1}(x_1)},$$

or in general, if $f_{X_j}(x_j) > 0$, then

$$f_{X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_n|X_j}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n|x_j) = \frac{f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)}{f_{X_j}(x_j)}.$$

We can also define

$$f_{X_1,X_3,X_5|X_2,X_4}(x_1, x_3, x_5|x_2, x_4) = \frac{f_{\boldsymbol{X}}(x_1, \ldots, x_n)}{f_{X_2,X_4}(x_2, x_4)}.$$

If it exists (i.e., if $f_{X_1}(x_1) > 0$ and the integral converges absolutely),

$$\mathrm{E}\left[g(X_2, \ldots, X_n)|x_1\right] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_2, \ldots, x_n) f_{X_2,\ldots,X_n|X_1}(x_s, \ldots, x_n|x_1)\ dx_2 \cdots dx_n.$$

Example 7.2.2. This example is from Parzen (1960, p. 340). Consider the decay of particles in a cloud chamber (or, similarly, the breakdown of equipment or the occurrence of accidents).

Assume that the time X of any particular particle to decay is a rv obeying an exponential probability law with parameter y. However, it is not assumed that the value of y is the same for all particles. Rather, it is assumed that there are particles of different types (or equipment of different types or individuals of different accident proneness). More specifically, it is assumed that for a particle randomly selected from the cloud chamber, the parameter y is a particular value of a random value obeying a Gamma probability law with pdf:

$$f_Y(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, \quad y > 0,$$

where α and β are positive constants characterizing the experimental conditions under which the particles are observed.

The assumption that the time X of a particle to decay obeys an exponential law is

$$f_{X|Y}(x|y) = ye^{-xy}, \quad x > 0.$$

Since $f_Y(y)$ and $f_{X|Y}(x|y)$ are assumed known, we can compute $f_{X,Y}(x,y) = f_{X|Y}(x|y) f_Y(y)$ and hence find the marginal law for the time to decay X (of a particle selected at random):

$$\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy \\
&= \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y)\, dy \\
&= \int_0^{\infty} ye^{-xy} \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}\, dy \\
&= \int_0^{\infty} \frac{\beta^\alpha}{\Gamma(\alpha)} y^\alpha e^{-y(\beta+x)}\, dy \\
&= \frac{\alpha\beta^\alpha}{(\beta+x)^{\alpha+1}} \underbrace{\int_0^{\infty} \frac{1}{\alpha\Gamma(\alpha)} (\beta+x)^{\alpha+1} y^\alpha e^{-y(\beta+x)}\, dy}_{=1} \\
&= \frac{\alpha\beta^\alpha}{(\beta+x)^{\alpha+1}}.
\end{aligned}$$

□

7.3 Joint Moment Generating Functions

Definition 7.3.1. If $f_{X,Y}(x,y)$ is the joint pdf/pmf of X and Y and if $\mathrm{E}\left[e^{t_1 X + t_2 Y}\right]$ exists for $t_1 \in (-h_1, h_1)$ and $t_2 \in (-h_2, h_2)$ for some $h_1, h_2 > 0$, then

$$M_{X,Y}(t_1, t_2) = \mathrm{E}\left[e^{t_1 X + t_2 Y}\right]$$

is called the *moment generating function* for the joint distribution of X and Y.

Example 7.3.1. Let X and Y be continuous rvs with joint pdf

$$f_{X,Y}(x,y) = e^{-y}, \quad 0 < x < y < \infty.$$

Thus,

$$\begin{aligned}
M_{X,Y}(t_1,t_2) &= \int_0^\infty \int_x^\infty e^{t_1x+t_2y}e^{-y}\ dydx \\
&= \int_0^\infty \int_x^\infty e^{t_1x}e^{y(t_2-1)}\ dydx \\
&= \int_0^\infty e^{t_1x} \int_x^\infty e^{y(t_2-1)}\ dydx \\
&= \frac{1}{t_2-1}\int_0^\infty e^{t_1x}\left[e^{y(t_2-1)}\right]_x^\infty\ dx \\
&= -\frac{1}{t_2-1}\int_0^\infty e^{t_1x+t_2x-x}\ dx \\
&= \frac{1}{1-t_2-1}\int_0^\infty e^{-x(1-t_1-t_2)}\ dx \\
&= \frac{1}{(1-t_2)(1-t_1-t_2)}\underbrace{\int_0^\infty (1-t_1-t_2)e^{-x(1-t_1-t_2)}\ dx}_{=1} \qquad \text{pdf for Exp}(\lambda = 1-t_1-t_2) \text{ distribution} \\
&= \frac{1}{(1-t_2)(1-t_1-t_2)}.
\end{aligned}$$

□

As with the univariate case, $M_{X,Y}(t_1,t_2)$ completely determined the joint distribution of X and Y. Note that

$$M_{X,Y}(t_1,0) = \mathrm{E}_{X,Y}\left[e^{t_1X}\right] = \int_{-\infty}^\infty \int_{-\infty}^\infty e^{t_1x} f_{X,Y}(x,y)\ dydx = \int_{-\infty}^\infty e^{t_1x} f_X(x)\ dx = M_X(t_1).$$

Similarly, $M_{X,Y}(0,t_2) = M_Y(t_2)$. Therefore,

$$\left.\frac{\partial M_{X,Y}(t_1,0)}{\partial t_1}\right|_{t_1=0} = \left.\frac{\partial M_X(t_1)}{\partial t_1}\right|_{t_1=0} = \mathrm{E}[X] \quad \Rightarrow \quad \frac{\partial M_{X,Y}(0,0)}{\partial t_1} = \mathrm{E}[X].$$

Similarly, $\frac{\partial M_{X,Y}(0,0)}{\partial t_2} = \mathrm{E}[Y]$. By the same reasoning,

$$\mathrm{E}\left[X^2\right] = \frac{\partial^2 M_{X,Y}(0,0)}{\partial t_1^2} \quad \Rightarrow \quad \mathrm{Var}(X) = \frac{\partial^2 M_{X,Y}(0,0)}{\partial t_1^2} - \left(\frac{\partial M_{X,Y}(0,0)}{\partial t_1}\right)^2$$

and analogously for $\mathrm{E}\left[Y^2\right]$ and $\mathrm{Var}(Y)$. The same logic can also be used to show

$$\mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right] = \frac{\partial^2 M_{X,Y}(0,0)}{\partial t_1 \partial t_2} - \left(\frac{\partial M_{X,Y}(0,0)}{\partial t_1}\right)\left(\frac{\partial M_{X,Y}(0,0)}{\partial t_2}\right).$$

Note that in general,

$$\begin{aligned}\frac{\partial^{k+m} M_{X,Y}(0,0)}{\partial t_1^k \partial t_2^m} &= \left.\frac{\partial^{k+m} M_{X,Y}(t_1,t_2)}{\partial t_1^k \partial t_2^m}\right|_{\substack{t_1=0\\t_2=0}} \\ &= \left.\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k y^m e^{t_1 x + t_2 y} f_{X,Y}(x,y)\; dxdy\right|_{\substack{t_1=0\\t_2=0}} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k y^m f_{X,Y}(x,y)\; dxdy \\ &= \mathrm{E}\left[X^k Y^m\right].\end{aligned}$$

Example 7.3.2 (Continuation of Example 7.3.1). Earlier, we showed that $M_{X,Y}(t_1,t_2) = \frac{1}{(1-t_2)(1-t_1-t_2)}$. From this,

$$\begin{aligned} M_{X,Y}(0,t_2) = \frac{1}{(1-t_2)^2} \quad &\Rightarrow \quad \left.\frac{\partial M_{X,Y}(0,t_2)}{\partial t_2}\right|_{t_2=0} = \left.\frac{2}{(1-t_2)^3}\right|_{t_2=0} = 2 = \mathrm{E}[Y], \\ M_{X,Y}(t_1,0) = \frac{1}{1-t_1} \quad &\Rightarrow \quad \left.\frac{\partial M_{X,Y}(t_1,0)}{\partial t_2}\right|_{t_1=0} = \left.\frac{1}{(1-t_1)^2}\right|_{t_1=0} = 1 = \mathrm{E}[X]. \end{aligned}$$

Furthermore,

$$\frac{\partial^2 M_{X,Y}(t_1,t_2)}{\partial t_1 \partial t_2} = 2(1-t_1-t_2)^{-3}(1-t_2)^{-1} + (1-t_1-t_2)^{-2}(1-t_2)^{-2} \quad \Rightarrow \quad \left.\frac{\partial^2 M_{X,Y}(t_1,t_2)}{\partial t_1 \partial t_2}\right|_{\substack{t_1=0\\t_2=0}} = 2+1 = 3 = \mathrm{E}[XY].$$

Therefore,

$$\mathrm{Cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] = 3 - 2 = 1. \qquad \square$$

Theorem 7.3.1. If the rvs X_1 and X_2 have mgf $M_{X_1,X_2}(t_1,t_2)$, then X_1 and X_2 are independent iff

$$M_{X_1,X-2}(t_1,t_2) = M_{X_1,X_2}(t_1,0) \cdot M_{X_1,X_2}(0,t_2) = M_X(t_1)M_Y(t_2).$$

Proof: Homework problem! ■

Example 7.3.3 (Continuation of Example 7.3.2).

$$M_{X,Y}(t_1,t_2) = \frac{1}{(1-t-1-t_2)(1-t_2)}, \quad M_{X,Y}(0,t_2) = \frac{1}{(1-t_2)^2}, \quad M_{X,Y}(t_1,0) = \frac{1}{1-t_1}.$$

Therefore, X and Y are not independent. $\square$

7.4 Order Statistics

Let $X_1, X_2, \ldots, X_n$ be iid from a continuous distribution with pdf $f_X(x)$. Since they are iid, their joint distribution is

$$\begin{aligned} f_{X_1X_2\cdots X_n}(x_1, x_2, \ldots, x_n) &= f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n) && \text{by independence} \\ &= f_X(x_1)f_X(x_2)\cdots f_X(x_n) && \text{identical distribution} \\ &= \prod_{i=1}^{n} f_X(x_i) \end{aligned}$$

Now let us define the random variables $Y_1, Y_2, \ldots, Y_n$, where $Y_i =$ the ith smallest value of $\{X_1, X_2, \ldots, X_n\}$. Thus, $Y_1 < Y_2 < \cdots < Y_n$. Note that

- Y_i is commonly denoted as $X_{(i)}$.
- Since this is a continuous distribution, $\mathrm{P}(Y_i = a) = 0$ for any value a. Therefore, for $i \neq j$,

$$\mathrm{P}(Y_i = Y_j) = \mathrm{P}([Y_i = a] \cap [Y_j = a] \text{ for some } a) = \int_{-\infty}^{\infty} \underbrace{\mathrm{P}(Y_i = a)\mathrm{P}(Y_j = a)}_{=0}\, da = 0.$$

As an example,

$$\begin{array}{ccc} X_1 = 10 & & Y_1 = 2 \\ X_2 = 3 & & Y_2 = 3 \\ X_3 = 7 & \Rightarrow & Y_3 = 7 \\ X_4 = 2 & & Y_4 = 8 \\ X_5 = 8 & & Y_5 = 10 \end{array}$$

However, note that we also have

$$\begin{array}{ccc} X_1 = 7 & & Y_1 = 2 \\ X_2 = 8 & & Y_2 = 3 \\ X_3 = 2 & \Rightarrow & Y_3 = 7 \\ X_4 = 10 & & Y_4 = 8 \\ X_5 = 3 & & Y_5 = 10 \end{array}$$

Generally, any permutation of the values of $(X_1, X_2, \ldots, X_n)$ is mapped to the same $(Y_1, Y_2, \ldots, Y_n)$. Since there are $n!$ such permutations,

$$\begin{aligned} f_{Y_1,Y_2,\ldots,Y_n}(y_1, y_2, \ldots, y_n) &= n! \cdot (\text{pdf of any one of those permutations}) \\ &= n! \cdot (\text{pdf of } X_1 = y_1,\ X_2 = y_2,\ \ldots,\ X_n = y_n) && \text{one such permutation} \\ &= n! \cdot f_{X_1X_2\cdots X_n}(y_1, y_2, \ldots, y_n) \\ &= n! \prod_{i=1}^{n} f_X(y_i) \end{aligned}$$

However, it is more useful to have the pdf for *one* of the order statistics, rather than the above formula for all of them at once.

Example 7.4.1. Let $X_1, X_2, X_3 \overset{iid}{\sim} \text{Exp}(1)$. Then $f_{X_i}(x_i) = e^{-x_i}, \quad x_i > 0$. For $Y_i = X_{(i)}$,

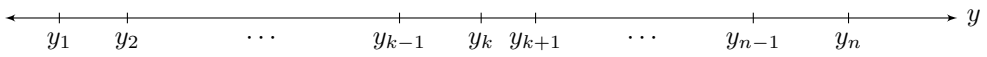
$$\begin{aligned}
f_{Y_1}(y_1) &= \int_{y_1}^{\infty}\int_{y_1}^{y_3} f_{Y_1,Y_2,Y_3}(y_1,y_2,y_3)\ dy_2 dy_3 \qquad \text{since } 0 < y_1 < y_2 < y_3 < \infty \\
&= \int_{y_1}^{\infty}\int_{y_1}^{y_3} 3! e^{-y_1}e^{-y_2}e^{-y_3}\ dy_2 dy_3 \\
&= 6e^{-y_1}\int_{y_1}^{\infty} e^{-y_3}\left[-e^{-y_2}\right]_{y_2=y_1}^{y_2=y_3}\ dy_3 \\
&= 6e^{-y_1}\int_{y_1}^{\infty} e^{-y_3}\left[-e^{-y_3}+e^{-y_1}\right]\ dy_3 \\
&= 6e^{-y_1}\int_{y_1}^{\infty} -e^{-2y_3}+e^{-y_1-y_3}\ dy_3 \\
&= 6e^{-y_1}\left[\frac{1}{2}e^{-2y_3}-e^{-y_1-y_3}\right]_{y_3=y_1}^{y_3=\infty} \\
&= 6e^{-y_1}\left(-\frac{1}{2}e^{-2y_1}+e^{-2y_1}\right) \\
&= 3e^{-3y_1}, \quad y_1 > 0.
\end{aligned}$$

□

For a given $k \in \{1, 2, \ldots, n\}$,

$$\begin{aligned}
f_{Y_k}(y_k) &= \int_{-\infty}^{y_k}\int_{y_1}^{y_3}\int_{y_2}^{y_4}\cdots\int_{y_{k-2}}^{y_k}\int_{y_k}^{y_{k+2}}\cdots\int_{y_{n-3}}^{y_{n-1}}\int_{y_{n-2}}^{y_n}\int_{y_{n-1}}^{\infty} n!\prod_{i=1}^{n} f_X(y_i)\ dy_n dy_{n-1}\ldots dy_{k+1}dy_{k-1}\ldots dy_1 \\
&= \frac{n!}{(k-1)!(n-k)!}[F_X(y_k)]^{k-1}f_X(y_k)[1-F_X(y_k)]^{n-k} \\
&= \binom{n}{k-1,\ 1,\ n-k}[F_X(y_k)]^{k-1}f_X(y_k)[1-F_X(y_k)]^{n-k}
\end{aligned}$$

The limits on the integrals derive from the fact that $-\infty < y_1 < y_2 < \cdots < y_{k-1} < y_k < y_{k+1} < \cdots < y_n < \infty$:



Let's think about what this formula means:

$$\underbrace{f_{Y_k}(y_k)}_{\approx \frac{1}{\varepsilon}\mathrm{P}(Y_k \approx_\varepsilon y_k)} = \underbrace{\binom{n}{k-1,\ 1,\ n-k}}_{\substack{\text{choose} \\ k-1 \quad \text{values to be } < y_k, \\ 1 \quad \text{value to be } = y_k, \\ n-k \quad \text{values to be } > y_k}} \cdot \underbrace{[F_X(y_k)]^{k-1}}_{[\mathrm{P}(X<y_k)]^{k-1}} \cdot \underbrace{f_X(y_k)}_{\approx \frac{1}{\varepsilon}\mathrm{P}(X \approx_\varepsilon y_k)} \cdot \underbrace{[1-F_X(y_k)]^{n-k}}_{[\mathrm{P}(X>y_k)]^{n-k}} \tag{7.4.1}$$

Above, the notation $\mathrm{P}(X \approx_\varepsilon y_k)$ means the probability that X *is in an epsilon neighborhood of* y_k, which means that

$$|X - y_k| < \varepsilon \qquad \Leftrightarrow \qquad y_k - \varepsilon < X < y_k + \varepsilon \qquad \text{for some small } \varepsilon > 0.$$

This gets into an interpretation of $f_X(x)$ which might be new to you:

$$\varepsilon f_X(x) \approx \mathrm{P}(|X - y_k| < \varepsilon) = \mathrm{P}(y_k - \varepsilon < X < y_k + \varepsilon) \qquad \text{for some small } \varepsilon > 0.$$

Example 7.4.2. Suppose $X_1, \ldots, X_{10} \overset{iid}{\sim} \mathrm{Exp}(1)$, so that $f_X(x) = e^{-x}$. Then for $Y_4 = X_{(4)}$,

$$f_{Y_4}(y_4) = \frac{10!}{3!6!} \left(1 - e^{-y_4}\right)^3 e^{-y_4} \left(e^{-y_4}\right)^6 \qquad \text{for } y_4 > 0,$$

since for $X \sim \mathrm{Exp}(\lambda)$, $F_X(x) = \displaystyle\int_0^x \lambda e^{-\lambda t}\, dt = \left[-e^{-\lambda t}\right]_0^x = 1 - e^{-\lambda x}$. □

Note that if $X_1, \ldots, X_n$ are joint *discrete* distributions, then there is a nonzero probability that $X_i = X_j$ even when $i \neq j$. However, that's the only difference – everything else is the same. For the joint probability of all order statistics, nothing changes. But now let's look at one particular order statistic, $X_(k); 1 \leq k \leq n; k, n \in \mathbb{N}$. Then $f_{Y(k)}(yk) = \binom{n}{k-1,1,n-k} \cdot ([F_X(y_k)]\,(k-1)) \cdot f_X(y_k) \cdot ([1F_X(y_k) + f_X(y_k)]\,(n-k))$. The only difference is in the last term, where $f_X(y_k)$ is added. This is because there is the possibility that some of the $k + n^{\text{th}}$ order statistics are equal to the k^{th}.

One last formula, that of joint i^{th} and j^{th} order statistic:

$$\begin{aligned} f_{Y_i,Y_j}(y_i, y_j) &= \binom{n}{i-1, 1, j-i-1, 1, n-j} \cdot [F_X(y_i)]^{i-1} \cdot f_X(y_i) \cdot [F_X(y_j) - F_X(y_i)]^{j-i-1} \cdot f_X(y_j) \cdot [1 - F_X(y_j)]^{n-j} \\ &= \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \cdot [F_X(y_i)]^{i-1} \cdot [F_X(y_j) - F_X(y_i)]^{j-i-1} \cdot [1 - F_X(y_j)]^{n-j} \cdot f_X(y_i) \cdot f_X(y_j) \end{aligned}$$

for $y_i < y_j$.

8 Limiting Distributions

8.1 Approximation of the Binomial Probability Law by the Normal and Poisson

Theorem 8.1.1 (DeMoivre-LaPlace Theorem). The probability that a rv having a binomial distribution with parameters n and p will have an observed values between a and b inclusive, for any two integers a and b, is approximately given by

$$\sum_{k=a}^{b} \binom{n}{k} p^k q^{n-k} \approx \frac{1}{\sqrt{2\pi}} \int_{\frac{a-np-1/2}{\sqrt{npq}}}^{\frac{b-np+1/2}{\sqrt{npq}}} e^{-y^2/2}\, dy = \Phi\left(\frac{b-np+1/2}{\sqrt{npq}}\right) - \Phi\left(\frac{a-np-1/2}{\sqrt{npq}}\right)$$

Proof: See an advanced probability text, such as Parzen (1960). ■

8.1.1 Normal Approximation to the Binomial

Let $X \sim \mathrm{B}(n,p) \quad \Rightarrow \quad \mathrm{E}[X] = np,\ \mathrm{Var}(X) = npq$. Let $Y_n = \frac{X-np}{\sqrt{npq}}$, and let $Z = \lim_{n\to\infty} Y_n$. Then $Z \sim \mathrm{N}(0,1)$. In practice, this means that for large n ($n > 30$ usually suffices), Y_n is approximately $\mathrm{N}(0,1)$.

Example 8.1.1. Let $X \sim \mathrm{B}(10, \frac{1}{5})$. Then

$$\mathrm{E}[X] = 10 \cdot \frac{1}{5} = 2,\ \mathrm{Var}(X) = 10 \cdot \frac{1}{5} \cdot \frac{4}{5} = \frac{8}{5} \quad \Rightarrow \quad \sigma_X \approx 1.26.$$

We want to find $\mathrm{P}(1 \le X \le 3)$. The actual probability yields

$$\mathrm{P}(1 \le X \le 3) = \sum_{x=1}^{3} \binom{10}{x} \left(\frac{1}{5}\right)^x \left(\frac{4}{5}\right)^{5-x} \approx 0.2684 + 0.3020 + 0.2013 = 77.17\%.$$

The approximation (which requires a continuity correction factor) yields

$$\begin{aligned}
\mathrm{P}\left(1 - \frac{1}{2} \le X \le 3 + \frac{1}{2}\right) &= \mathrm{P}\left(0.5 \le X \le 3.5\right) \\
&= \mathrm{P}\left(\frac{0.5-2}{1.265} \le \frac{X-2}{1.265} \le \frac{3.5-2}{1.265}\right) \\
&= \mathrm{P}(-1.186 \le Z \le 1.186) \\
&= \Phi(1.186) - \Phi(-1.186) \\
&= \Phi(1.186) - [1 - \Phi(1.186)] \\
&= 2\Phi(1.186) - 1 \\
&\approx 2 \cdot 0.8830 - 1 \\
&= 76.60\%.
\end{aligned}$$

□

In general, for $a, b \in \mathbb{Z}^*$,

$$\mathrm{P}(a \le X \le b) = \mathrm{P}(a - \tfrac{1}{2} \le X \le b + \tfrac{1}{2}) = \mathrm{P}\left(\frac{a - \frac{1}{2} - np}{\sqrt{npq}} \le \frac{X - \frac{1}{2} - np}{\sqrt{npq}} \le \frac{b + \frac{1}{2} - np}{\sqrt{npq}} \right) \approx \mathrm{P}\left(\frac{a - \frac{1}{2} - np}{\sqrt{npq}} \le Z \le \frac{b + \frac{1}{2} - np}{\sqrt{npq}} \right).$$

Example 8.1.2. Binomial with $n = 100$, $p = 0.3$:

# Successes	Binomial Probability	Normal Approximation	% Error
$9 \le X \le 11$	0.000006	0.00003	+400%
$12 \le X \le 14$	0.00015	0.00033	+100%
$15 \le X \le 17$	0.00201	0.00283	+40%
$18 \le X \le 20$	0.1430	0.1599	+12%
$21 \le X \le 23$	0.05907	0.05895	0%
$24 \le X \le 26$	0.14887	0.14447	-3%
$27 \le X \le 29$	0.23794	0.23405	-2%
$31 \le X \le 33$	0.23013	0.23405	+2%
$34 \le X \le 36$	0.14086	0.14447	+3%
$37 \le X \le 39$	0.05889	0.05895	0%
$40 \le X \le 42$	0.01702	0.01599	-6%
$43 \le X \le 45$	0.00343	0.00283	-18%
$46 \le X \le 48$	0.00049	0.00033	-33%
$49 \le X \le 51$	0.00005	0.00003	-40%

□

Example 8.1.3. Toss a fair die $n = 6000$ times. The probability that a 3 will occur between 980 and 1030 times is

$$\begin{aligned}
\sum_{k=980}^{1030} \binom{6000}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{6000-k} &\approx \Phi\left(\frac{1030 - 1000 + \frac{1}{2}}{\sqrt{6000 \cdot \frac{1}{6} \cdot \frac{5}{6}}} \right) - \Phi\left(\frac{980 - 1000 - \frac{1}{2}}{\sqrt{6000 \cdot \frac{1}{6} \cdot \frac{5}{6}}} \right) \\
&\approx \Phi(1.06) - \Phi(-0.71) \\
&\approx 0.8554 - 0.2389 \\
&\approx 0.62.
\end{aligned}$$

□

Rule of Thumb: Use the normal approximation if $np(1-p) \ge 3$.

The Poisson approximation to the binomial is used when the binomial pmf is far from being bell-shaped – usually $p \le 0.1$ and n is large but np is not large:

$$\binom{n}{k} p^k q^{n-k} \approx e^{np} \frac{(np)^k}{k!}$$

This was proven in a previous chapter.

Historically, the Poisson approximation to the binomial might have been used because the term $\binom{n}{k}$ was difficult to compute for large n. However, with modern computers this is no longer true, so the Poisson approximation is hardly used. It is, however, useful to remember the role of the binomial (actually, Bernoulli) distribution in the derivation of a Poisson process, as explained in section 4.1.5.

When np becomes large, the Poisson is approximated by the normal (which will be proven by the Central Limit Theorem in the next section). Let $Y \sim \mathrm{P}(\lambda)$ with $\lambda = np$. Then

$$\frac{Y-\lambda}{\sqrt{\lambda}} = \frac{Y-np}{\sqrt{np}} \text{ is approximately } \mathrm{N}(0,1) \text{ for large } n.$$

Example 8.1.4. Comparison of the Poisson to the normal approximation: Values of

$$\mathrm{P}(a \le X \le b) = \sum_{x=1}^{b} e^{-100}\frac{100^x}{x!}$$

for $\lambda = 100$:

Quantity	Actual Probability	Normal Approximation	% Error
$85 \le X \le 90$	0.11384	0.11049	
$90 \le X \le 95$	0.18485	0.17950	
$95 \le X \le 105$	0.41763	0.41768	
$90 \le X \le 110$	0.70652	0.70628	
$110 \le X \le 115$	0.10738	0.11049	
$115 \le X \le 120$	0.05323	0.05335	

□

Example 8.1.5. From Parzen (1960): Suppose you are designing the physical premises of a newly-organized research lab. Since there will be a large number of private offices in the lab, there will also be a large number n of individual telephones, each connecting to a central telephone switchboard. the question arises: How many outside lines will the switchboard require to establish a fairly high probability – say, 95% – that any person who desires the use of an outside telephone line (whether on the outside of the lab calling in or on the inside calling out) will find one immediately available?

Regard the problem as one involving independent Bernoulli trials. We suppose that for each telephone in the lab, say the j^{th} phone, that there is a probability p_j that an outside line will be required (either for an incoming or outgoing call).

One could *estimate* p_j by observing in the course of an hour how many minutes h_j an outside line is engaged and estimating p_j as $\frac{h_j}{60}$. In order to have repeated Bernoulli trials, we assume that $p_1 = p_2 = \cdots Pp_n = p$.

We next assume that the event that the j^{th} phone requires an outside line is independent of the event that the k^{th} phone requires an outside line for $j, k \in \{1, 2, \ldots, n\}$. The problem is that exactly k outside lines will be in demand at a given moment is

$$\binom{n}{k} p^k q^{n-k}.$$

Let K denote the number of outside lines connected to the lab switchboard.

$$\mathrm{P}(\#\text{ lines in demand } \le K) = \sum_{k=0}^{K} \binom{n}{k} p^k q^{n-k} \ge 0.95.$$

If $0.1 \le p \le 0.9$ and n is large, then

$$\sum_{k=0}^{K} \binom{n}{k} p^k q^{n-k} \approx \Phi\left(\frac{K-np+\frac{1}{2}}{\sqrt{npq}}\right) - \Phi\left(\frac{K-np+\frac{1}{2}}{\sqrt{npq}}\right) \ge 0.95$$

If $p \leq 0.1$ and n is large (but np is not "too large"), then

$$\sum_{k=0}^{K} \binom{n}{k} p^k q^{n-k} \approx \sum_{k=0}^{K} \frac{e^{-np}(np)^k}{k!} \geq 0.95$$

In both of the above equations, we need to solve for K:

Value of K for those two equations

p		$\frac{1}{30}$		$\frac{1}{10}$		$\frac{1}{3}$	
Approximation		Poisson	Normal	Poisson	Normal	Poisson	Normal
P = 0.95	$n = 90$	6	5.3	14	13.2	39	36.9
	$n = 900$	39	38.4	106	104.3	–	322.8
P = 0.99	$n = 90$	8	6.5	17	15.1	43	39.9
	$n = 900$	43	42	113	110.4	–	332.4

□

8.2 Central Limit Theorem

We already know that there is a 1-1 relationship between $F_X(\cdot)$ and $\phi_X(\cdot)$. We will now state a theorem that further related distribution functions to characteristic functions.

Theorem 8.2.1 (Continuity Theorem). Let $\{X_n : n \in \mathbb{Z}^+\}$ and X be rvs such that

$$\lim_{n\to\infty} \phi_{X_n}(t) = \phi_X(t), \quad -\infty < t < \infty$$

Then

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

at all points x where $F_X(\cdot)$ is continuous.

Proof: Involved – see an advanced probability text. ■

The continuity theorem states that convergence of characteristic functions implies convergence of the corresponding distribution functions. In other words, distribution functions depend continuously on their characteristic functions.

For example, if $\lim_{n\to\infty} \phi_{X_n}(t) = e^{-t^2/2}$, then $\lim_{n\to\infty} F_{X_n}(x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy$.

Before we prove the Central Limit Theorem, we prove a needed lemma:

Lemma 8.2.1. Let X be an rv having a characteristic function $\phi_X(t)$, a finite mean $\mathrm{E}[X] = \mu$, and a finite variance σ^2. Then

$$\lim_{t\to 0} \frac{\ln \phi_X(t) - i\mu t}{t^2} = -\frac{\sigma^2}{2}.$$

Proof: $\phi_X(t)$ is continuous in t and $\phi_X(0) = 1$. Thus, $\ln \phi_X(t)$ is well defined for t near 0. Note that $\ln \phi_X(0) = 0$. We know that $\phi'_X(0) = i\mu$ and that $\phi''_X(0) = i^2 \mathrm{E}\left[X^2\right] = -\mathrm{E}\left[X^2\right] = -(\mu^2 + \sigma^2)$. Thus,

$$\begin{aligned}
\lim_{t\to 0} \frac{\ln \phi_X(t) - i\mu t}{t^2} &= \lim_{t\to 0} \frac{\phi'_X(t)/\phi_X(t) - i\mu}{2t} && \text{L'Hôpital's Rule} \\
&= \lim_{t\to 0} \frac{\phi'_X(t) - i\mu\phi_X(t)}{2t\phi_X(t)} \\
&= \lim_{t\to 0} \frac{\phi''_X(t) - i\mu\phi'_X(t)}{2t\phi'_X(t) + 2\phi_X(t)} && \text{L'Hôpital's Rule (again)} \\
&= \frac{-(\mu^2 + \sigma^2) - i\mu(i\mu)}{0 \cdot i\mu + 2 \cdot 1} \\
&= \frac{-\mu^2 - \sigma^2 + \mu^2}{2} \\
&= -\frac{\sigma^2}{2}.
\end{aligned}$$

■

Theorem 8.2.2 (Central Limit Theorem). Let $X_1, \ldots, X_n$ be iid, each having a finite mean μ and a finite variance σ^2. Let $S_n = X_1 + \cdots + X_n$. Then

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right) = \Phi(x), \quad x \in \mathbb{R}.$$

Proof: Let $S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}$. Note that S_n^* is the standardized rv corresponding to S_n. That is, $\mathrm{E}\left[S_n^*\right] = 0$ and $\mathrm{Var}\left(S_n^*\right) = 1$.

$$\begin{aligned}
\phi_{S_n^*}(t) &= \exp\left(\frac{-in\mu t}{\sigma\sqrt{n}}\right) \cdot \phi_{S_n}\left(\frac{t}{\sigma\sqrt{n}}\right) \\
&= \exp\left(\frac{-in\mu t}{\sigma\sqrt{n}}\right) \cdot \phi_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right)^n \\
&= \exp\left(\frac{-in\mu t}{\sigma\sqrt{n}}\right) \cdot \exp\left[n \ln \phi_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right)\right] \\
&= \exp\left[n \ln \phi_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right) - \frac{in\mu t}{\sigma\sqrt{n}}\right].
\end{aligned}$$

We will show that

$$\lim_{n\to\infty} \left[n \ln \phi_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right) - \frac{in\mu t}{\sigma\sqrt{n}}\right] = \frac{-t^2}{2} \tag{8.2.1}$$

Note that (8.2.1) is true if $t = 0$. If $t \neq 0$, then the left side of (8.2.1) may be written as

$$\frac{t^2}{\sigma^2} \lim_{n\to\infty} \left[\frac{\ln \phi_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right) - i\mu\left(\frac{t}{\sigma\sqrt{n}}\right)}{\left(\frac{t}{\sigma\sqrt{n}}\right)^2}\right] \tag{8.2.2}$$

by multiplying (8.2.1) by $\frac{t^2/\sigma^2}{t^2/\sigma^2}$. In (8.2.2), when $n \to \infty$, $\frac{t}{\sigma\sqrt{n}} \to 0$. Thus, we can write (8.2.2) as

$$\frac{t^2}{\sigma^2} \lim_{\frac{t}{\sigma\sqrt{n}} \to 0} \left[\frac{\ln \phi_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right) - i\mu\left(\frac{t}{\sigma\sqrt{n}}\right)}{\left(\frac{t}{\sigma\sqrt{n}}\right)^2} \right] \underbrace{=}_{\text{Lemma 8.2.1}} \frac{t^2}{\sigma^2}\left(\frac{-\sigma^2}{2}\right) = \frac{-t^2}{2}.$$

Therefore, we have proven (8.2.1), and it thus follows that

$$\lim_{n\to\infty} \phi_{S_n^*}(t) = \lim_{n\to\infty} \exp\left[n\phi_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right) - \frac{in\mu t}{\sigma\sqrt{n}} \right] = \exp\left[\lim_{n\to\infty} \frac{t}{\sigma\sqrt{n}} - \frac{in\mu t}{\sigma\sqrt{n}} \right] = \exp\left(\frac{-t^2}{2}\right).$$

But this is the characteristic function for a $\mathrm{N}(0,1)$ rv. Therefore, by the continuity theorem,

$$\lim_{n\to\infty} \mathrm{P}\left(S_n^* \le x\right) = \Phi(x).$$

■

By the CLT, we obtain that for large n and $S_n = X_1 + \cdots + X_n$, where X_i are iid rvs,

$$\mathrm{P}(S_n \le x) = \mathrm{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le \frac{x - n\mu}{\sigma\sqrt{n}}\right) \approx \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right).$$

Example 8.2.1. Suppose that the life length of a certain kind of light bulb, after it is installed, is exponentially distributed with mean life length of 10 days. As soon as one light burns out, a similar one is installed. Find the probability that more than 50 bulbs will be required during a one-year period.

Let X_n denote the life length of the n^{th} light bulb that is installed.

Assume that $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp(mean = 10 days)} = \text{Exp}\left(\lambda = \frac{1}{10 \text{ days}}\right)$. S_n denotes the time when the n^{th} bulb burns out. We want to find $\mathrm{P}(S_{50} < 365)$. Since

$$\mathrm{E}\left[S_{50}\right] = \frac{50}{\lambda} = 500, \quad \mathrm{Var}\left(S_{50}\right) = \frac{50}{\lambda^2} = 5000,$$

then the CLT tells us that

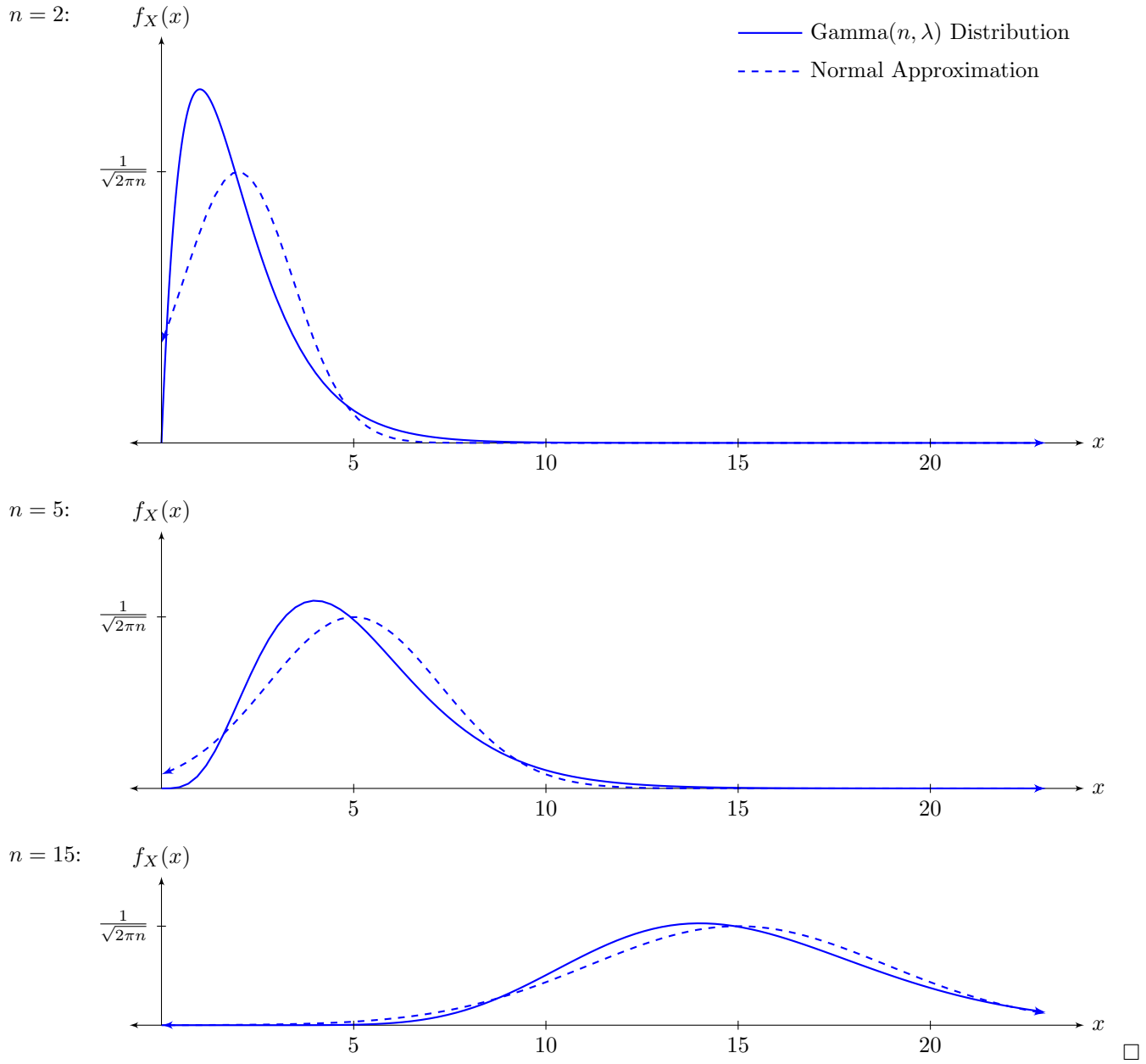
$$\mathrm{P}\left(S_{50} < 365\right) \approx \Phi\left(\frac{365 - 500}{\sqrt{5000}}\right) \approx \Phi(-1.91) \approx 0.028.$$

Thus, it is unlikely that more than 50 bulbs will be needed. □

Example 8.2.2. Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(1)$. It can be shown that $S_n = X_1 + \cdots = X_n \sim \text{Gamma}(n, \lambda)$. However, for large values of n,

$$\text{P}(S_n \leq x) = \text{P}\left(\frac{S_n - n}{\sqrt{n}} \leq \frac{x-n}{\sqrt{n}}\right) \approx \Phi\left(\frac{x-n}{\sqrt{n}}\right).$$

In other words, the distribution of S_n is approximately $\text{N}(n, \sigma^2 = n)$. This can be shown graphically.



□

A more general form of the CLT can be proven. If $X_1, X_2, \ldots$ are independent rvs with finite means and finite variances, and if $\forall k$, $\frac{\sigma_k}{\sigma[S_n]} < \varepsilon \ \forall \varepsilon > 0$, and if n is sufficiently large, then

$$\frac{S_n - \text{E}[S_n]}{\sigma[S_n]} \to \text{N}(0, 1).$$

Example 8.2.3. Candidates A and B are running for office and 45% of the electorate favor candidate A. What is the probability that in a sample of size 100, at least 50% of those samples will favor candidate A?

Let

$$X_i = \begin{cases} 1 & \text{if person } i \text{ favors candidate } A \\ 0 & \text{otherwise} \end{cases}.$$

Then $X_1, \ldots, X_{100}$ are independent Bernoulli rvs where $p = \mathrm{P}(X_i = 1) = 0.45$. Then $\sum_{i=1}^{100} X_i$ is the number of people sampled who will favor A, and $\sum_{i=1}^{100} X_i \sim \mathrm{B}(100, 0.45)$. We want to find $\mathrm{P}\left(\sum_{i=1}^{100} X_i \geq 50\right)$.

We could do this via a computer or with the Binomial tables. However, it is easier to use the CLT, which tells us that

$$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} \dot{\sim} \mathrm{N}(0,1) \quad \text{for large } n.$$

Thus,

$$\mathrm{P}\left(\sum_{i=1}^{100} X_i \geq 50\right) = \mathrm{P}\left(\frac{\sum_{i=1}^{100} X_i - 100(0.45)}{\sqrt{100(0.45)(0.55)}} \geq \frac{50 - 100(0.45) + \frac{1}{2}}{\sqrt{100(0.45)(0.55)}}\right) \approx \mathrm{P}\left(Z \geq 1.11\right) = \Phi(1.11) \approx 13\%. \qquad \square$$

Example 8.2.4. If each strand in a rope has a breaking strength with mean 20 pounds and sd 2 pounds and the breaking strength of a rope is the sum of the (independent) breaking strengths of all the strands, what is the probability that a rope made up of 64 strands will support a weight of 1280 pounds or more?

Let the rv X denote the breaking strength of the i^{th} strand. Given that $\mathrm{E}[X_i] = 120$, $\sigma[X_i] = 2$ $\forall i$, and that $X_1, \ldots, X_{64}$ are all independent, we know by the CLT that

$$\frac{\sum_{i=1}^{n} X_i - 1280}{2\sqrt{64}} \dot{\sim} \mathrm{N}(0,1) \quad \text{for large } n.$$

Thus,

$$\mathrm{P}\left(\sum_{i=1}^{64} X_i > 1280\right) = \mathrm{P}\left(\frac{\sum_{i=1}^{64} X_i - 64(20)}{2\sqrt{64}} > \frac{1280 - 1280}{2}\sqrt{64}\right) \approx \mathrm{P}\left(Z \geq 0\right) = \Phi(0) = 50\%.$$

If we want to calculate the probability that the rope will support a weight of 1240 pounds, then

$$\mathrm{P}\left(\sum_{i=1}^{64} X_i > 1240\right) = \mathrm{P}\left(\frac{\sum_{i=1}^{64} X_i - 64(20)}{2\sqrt{64}} > \frac{1240 - 1280}{2}\sqrt{64}\right) \approx \mathrm{P}\left(Z \geq -\frac{40}{16}\right) = \Phi(-2.5) = 99\%. \qquad \square$$

8.3 Chebyshev's Inequality

Suppose that the mean and variance of an rv X are known. We would like to calculate $\mathrm{P}(\mu - h\sigma \le X \le \mu + h\sigma)$, where $\mu = \mathrm{E}[X]$, $\sigma = \sigma[X]$, and $h \ge 0$.

- If we know the functional form of the pmf or pdf, then it may be possible to relate the unknown parameters to the mean, μ, and the variance σ^2. As examples,
  - If $X \sim \mathrm{N}(\mu, \sigma^2)$, then

$$\mathrm{P}\left(\mu - h\sigma \le X \le \mu + h\sigma\right) = \mathrm{P}\left(-h \le \tfrac{X-\mu}{\sigma} \le h\right) = \Phi(h) - \Phi(-h) = \int_{-h}^{h} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy.$$

  - If $X \sim \mathrm{Exp}\left(\mu = \frac{1}{\lambda}\right)$, then $\mu = \frac{1}{\lambda}$ and $\sigma^2 = \frac{1}{\lambda^2}$, and

$$\begin{aligned}\mathrm{P}\left(\mu - h\sigma \le X \le \mu + h\sigma\right) &= \mathrm{P}\left(\tfrac{1-h}{\lambda} \le X \le \tfrac{1+h}{\lambda}\right) \\ &= \begin{cases} \displaystyle\int_{\frac{1-h}{\lambda}}^{\frac{1+h}{\lambda}} \lambda e^{-\lambda y}\, dy = \left[-e^{-\lambda y}\right]_{\frac{1-h}{\lambda}}^{\frac{1+h}{\lambda}} = e^{-1+h} - e^{-1-h} & h < 1 \\ \displaystyle\int_{0}^{\frac{1+h}{\lambda}} \lambda e^{-\lambda y}\, dy = \left[-e^{-\lambda y}\right]_{0}^{\frac{1+h}{\lambda}} = 1 - e^{-1-h} & h \ge 1 \end{cases}.\end{aligned}$$

- If we don't know the functional form of the pmf or pdf, then a crude estimate of $\mathrm{P}(\mu - h\sigma \le X \le \mu + h\sigma)$ can be obtained.

A Russian probabilist, Chebyshev, with the help of another Russian, Markov, found a lower bound for this probability using only the information that the rv X has a finite mean μ and variance σ^2.

Theorem 8.3.1 (Markov's Inequality). Let the rv X have the property that $\mathrm{P}(X < 0) = 0$ and $\mathrm{E}[X] = a < \infty$, where $a > 0$. Then for any $k \ge 1$,

$$\mathrm{P}(X < ak) \ge 1 - \frac{1}{k}.$$

Proof: The proof will be for the continuous case – the discrete case is similar.

By hypothesis,

$$\int_{-\infty}^{0} f_X(x)\, dx = 0 \quad \Rightarrow \quad \int_{-\infty}^{0} x f_X(x)\, dx = 0.$$

Thus,

$$a = \mathrm{E}[X] = \int_{0}^{\infty} x f_X(x)\, dx \ge \int_{ak}^{\infty} x f_X(x)\, dx \ge ak \int_{ak}^{\infty} f_X(x)\, dx = ak(1 - \mathrm{P}(X < ak).$$

Dividing both sides by $ak > 0$ gives

$$\frac{1}{k} \ge 1 - \mathrm{P}(X < ak) \quad \Leftrightarrow \quad 1 - \frac{1}{k} \le \mathrm{P}(X < ak).$$

■

Example 8.3.1. Let X be a discrete rv s.t., $\mathrm{P}(X \leq 0) = 0$ and $\mathrm{E}[X] = 5$. Then

$$\text{if } k = 2, \text{ then } \mathrm{P}(X < 10) = \mathrm{P}(X < 5 \cdot 2) \geq 1 - \frac{1}{2} = \frac{1}{2}$$
$$\text{if } k = \tfrac{7}{5}, \text{ then } \mathrm{P}(X < 7) = \mathrm{P}(X < 5 \cdot \tfrac{7}{5}) \geq 1 - \frac{5}{7} = \frac{2}{7},$$
$$\text{if } k = 1, \text{ then } \mathrm{P}(X < 5) = \mathrm{P}(X < 5 \cdot 1) \geq 1 - \frac{1}{1} = 0.$$

Of course, that last equation is not very helpful – again, this is a very rough lower bound. However, note that this lower bound is indeed attainable. If $\mathrm{P}(X = 5) = 1$, then we still have $\mathrm{E}[X] = 5$, and $\mathrm{P}(X < 5) = 0$. □

Theorem 8.3.2 (Chebyshev's Inequality). If the rv Y has nonzero finite variance σ^2, then $\forall h \geq 0$,

$$\mathrm{P}\left(|Y - \mu| < h\sigma\right) \geq 1 - \frac{1}{h^2},$$

where $\mu = \mathrm{E}[Y]$.

Proof: If $h < 1$, the inequality is true, since $1 - \frac{1}{h^2} < 0$. Let $h \geq 1$ be given. Let $X = (Y - \mu)^2$. Then $\mathrm{P}(X < 0) = 0$ and $\mathrm{E}[X] = \sigma^2 \in (0, \infty)$ by hypothesis. From Markov's Inequality with $a = \sigma^2$ and $k = h^2$, we have

$$\mathrm{P}\left(|Y - \mu| < h\sigma^2\right) = \mathrm{P}\left((Y - \mu)^2 < h^2\sigma^2\right) = \mathrm{P}(X < ka) \geq 1 - \frac{1}{k} = 1 - \frac{1}{h^2}.$$

■

Under the general assumptions of the theorem, only that Y has a finite and non-zero variance, the inequality cannot be improved, as shown in Example 8.3.2.

Example 8.3.2. This is from Birnbaum (1962, pp. 87-88). Let $x_0 \in \mathbb{R}$, and let $t, \sigma \in \mathbb{R}$ be given s.t. $t > 1$, $\sigma > 0$. Then let the rv X have possible values

$$x_0 - t\sigma,\ x_0,\ x_0 + t\sigma$$

with probabilities

$$\frac{1}{2t^2},\ 1 - \frac{1}{t^2},\ \frac{1}{2t^2}.$$

Then X has expectation x_0 and variance σ^2. Furthermore,

$$\mathrm{P}\left(|X - x_0| < t\sigma\right) = \mathrm{P}(X - x_0) = 1 - \frac{1}{t^2},$$

which is the lower bound shown by Chebyshev's Inequality. Here, Chebychev's Inequality gives a lower bound for the probability that X will fall into an interval of width $2t\sigma$, with the midpoint x_0. □

Chebyshev's inequality states that with $h = 4$, the probability is at least 93.75% that an observed value of X will lie within 4 standard deviations of the mean. That is, that

$$\mathrm{P}(-4\sigma < y - \mu_Y < 4\sigma) \geq 1 - \frac{1}{16} = 93.75\%.$$

We know that is $X \sim \mathrm{N}(\mu, \sigma)$, then

$$\mathrm{P}(-2.58\sigma < X - \mu_X < 2.58\sigma) \approx 99\%.$$

Chebyshev's inequality may be restated in the form

$$\mathrm{P}\left(|y - \mu_Y| \geq h\sigma\right) < \frac{1}{h^2}.$$

8.4 Laws of Large Numbers

Theorem 8.4.1 (Weak Law of Large Numbers). Let X be an rv with finite variance σ^2 and mean μ. Let $X_1, \ldots, X_n$ be a countable sequence of independent repetitions of X. then

$$\mathrm{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| < \varepsilon\right) \geq 1 - \frac{\sigma^2}{n\varepsilon^2} \quad \forall \varepsilon > 0. \tag{8.4.3}$$

Proof: Let $\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$. Then

$$\mathrm{E}\left[\bar{X}_n\right] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[X_i\right] = \frac{1}{n} \cdot n\mathrm{E}[X] = \mu.$$
$$\sigma^2_{\bar{X}_n} = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

Assume $\sigma^2 > 0$. Then

$$\begin{aligned}
\mathrm{P}\left(\left|\bar{X}_n - \mu\right| < \varepsilon\right) &= \mathrm{P}\left(\left|\bar{X}_n - \mu\right| < \sigma_{\bar{X}_n} \cdot \frac{\varepsilon}{\sigma_{\bar{X}_n}}\right) \\
&= \mathrm{P}\left(\left|\bar{X}_n - \mu\right| < h\sigma_{\bar{X}_n}\right) && \text{where } h = \tfrac{\varepsilon}{\sigma_{\bar{X}_n}} \\
&\geq 1 - \frac{1}{h^2} && \text{from Chebychev's inequality} \\
&= 1 - \frac{1}{\varepsilon^2/\sigma^2_{\bar{X}_n}} \\
&= 1 - \frac{\sigma^2_{\bar{X}_n}}{\varepsilon^2} \\
&= 1 - \frac{\sigma^2}{n\varepsilon^2}.
\end{aligned}$$

Note that is $\sigma^2 = 0$, then $\mathrm{P}(|\bar{X}_n - \mu| < \varepsilon) = 1\ \forall \varepsilon > 0$. ■

Example 8.4.1. Assume that $\mathrm{Var}(X) = 2$.[1] The mean and the functional form of the pdf/pmf is not known. How large a sample should one take in order to have $\mathrm{P}\left(\left|\bar{X}_n - \mu\right| < 0.02\right) \geq 95\%$?

We want to find n in the WLLN. Assume that $\sigma^2 = 2,\ \varepsilon = 0.02$. Then

$$\begin{aligned}
\mathrm{P}\left(\left|\bar{X}_n - \mu\right| < 0.02\right) \geq 1 - \frac{\sigma^2}{n\varepsilon^2} \geq 0.95 \quad &\Rightarrow \quad \frac{\sigma^2}{n\varepsilon^2} \leq 0.05 \\
&\Rightarrow \quad n \geq \frac{\sigma^2}{0.05\varepsilon^2} = \frac{2}{0.05 \cdot 0.04^2} = 100,000.
\end{aligned}$$

What approximation would you use if you knew μ and σ^2? □

[1] Actually, we can assume that $\mathrm{Var}(X) \leq 2$ – this is just an upper bound for $\mathrm{Var}(X)$

Example 8.4.2. How many trials on an experiment with two outcomes (A and B) should be performed in order that the probability will be 95% or greater that the observed relative frequency of A will differ from the probability p of the occurrence of A by no more than 0.02?

We will use the law of large numbers. Let

$$X = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise (if } B \text{ occurs)} \end{cases}.$$

The observed relative frequency of A is $\bar{X}_n = \frac{X_1+\cdots+X_n}{n}$. $\mathrm{P}(A) = \mathrm{P}(X_i = 1) = p$ and $\mathrm{E}[X_i] = \mathrm{E}\left[\bar{X}_n\right] = p$. We want to find $n \ni \mathrm{P}\left(\left|\bar{X}_n - p\right| < \varepsilon\right) \geq 0.95$, where $\varepsilon = 0.02$, i.e., $\frac{\sigma^2}{n\varepsilon^2} \leq 0.05$.

We know that $\sigma_X^2 = p(1-p) \leq \frac{1}{4}$, which we can show by

$$p(1-p) - \frac{1}{4} = p - p^2 - \frac{1}{4} = -\left(p^2 - p + \frac{1}{4}\right) = -\left(p - \frac{1}{2}\right)^2 \leq 0$$

and adding $\frac{1}{4}$ to both sides. Therefore,

$$\frac{\sigma^2}{n\varepsilon^2} \leq \frac{1/4}{n\varepsilon^2} \leq 0.05 \quad \Rightarrow \quad n \geq \frac{1}{4 \cdot 0.02^2 \cdot 0.05} \approx 12,500.$$

Note that if we knew p, we would use the fact that $\frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}} \dot{\sim} \mathrm{N}(0,1)$. □

Corollary 8.4.1. If the hypothesis of the WLLN are satisfied, then

$$\lim_{n\to\infty} \mathrm{P}\left(\left|\bar{X}_n - \mu\right| < \varepsilon\right) = 1 \quad \forall \varepsilon > 0. \tag{8.4.4}$$

Proof: This follows immediately from Equation (8.4.3). ■

Note that one can prove (8.4.4) using only the assumption that $\mathrm{E}[X_i] = \mu$ is finite. That is, one does not need to assume a finite variance for X_i.

9 Markov Chains

9.1 Mathematical Model

Consider a physical system such as emissions from a radioactive source. Assume that this system is observed periodically and that the state of the system at each observation is noted. For emissions, the state of the system at each time t is the total number of emissions that have occurred during the time interval from 0 to t.

Let the possible states of the system be denoted $E_0, E_1, E_2, \ldots$. We will use the variable X_i to represent the outcome of the i^{th} observation of the system. Thus, X_i can take as a value any one of the states $E_0, E_1, E_2, \ldots$. After n observations of the system, we have a sample $(X_1, \ldots, X_n)$. For radioactive emissions, the possible states are $\{E_i = i \text{ emissions} : \ i = 0, 1, \ldots, m\}$.

If the system is observed every 10 seconds and we start at time 10 seconds, then the i^{th} observation is taken at time $10i$ seconds. On the i^{th} observation, we observe one of the possible states.

A possible sample after eight observations is $(X_1, \ldots, X_8) = (E_2, E_4, E_4, E_5, E_8, E_9, E_9, E_{10})$. Thus, on the 5^{th} observation (at time 50 seconds), state E_8 (total of 8 emissions) was observed.

Since the sample is obtained sequentially, an important question that can be asked is:

Does our knowledge of past history of the system affect the probability of future events?

For example, does knowledge of the outcome on the first $k-1$ observations affect our probability of some particular state, say E_i, on the k^{th} observation?

Obviously for radioactive emissions, knowledge of the particular state observed on the $k+1^{\text{st}}$ observation affects our probability of observing E_i on the k^{th} observation. E.g., knowing that state E_{10} (total of 10 emissions) was observed on the 8^{th} observation affects the probability of observing E_2 (total of 2 emissions) on the next (9^{th}) observation (i.e., this has a probability of zero).

We can restate the general question again in the following form: What is the probability of observing state E_i on the k^{th} observation, knowing the particular states that were observed at each of the previous $k-1$ observations? This probability may be expressed as

$$\mathrm{P}\left(X_k = E_i | X_1 = E_{j1}, X_2 = E_{j2}, \ldots, X_{k-1} = E_{jk-1}\right),$$

where E_{jn} for $n = 1, \ldots, k-1$ denotes the state that was observed on the n^{th} observation.

Assume the outcomes of the observations in a system are independent of one another (in our example they are dependent). It can be shown that

$$\mathrm{P}\left(X_k = E_i | X_1 = E_{j1}, X_2 = E_{j2}, \ldots, X_{k-1} = E_{jk-1}\right) = \mathrm{P}\left(X_k = E_i\right).$$

Example 9.1.1. A simple example of this type of system, where the observations are independent of one another, is that of tossing repeatedly a fair coin. The possible states are { heads, tails }. The probability of observing a head on the 4^{th} toss, given that the first 3 tossses were tails, is still the (unconditional) probability of observing a head on the 4^{th} toss – i.e., $\frac{1}{2}$. Using the notation defined above, where $E_1 = \{\text{head}\}$ and $E_2 = \{$ tail $\}$,

$$\mathrm{P}\left(X_4 = E_1 | X_1 = E_2, X_2 = E_2, X_3 = E_2\right) = \mathrm{P}(X_4 = E_1) = \tfrac{1}{2}. \qquad \square$$

In general, however, many physical systems show dependence, and the state that occurs on the k^{th} observation is conditioned by the particular states through which the system has passed before reaching the k^{th} state. For a probabilistic system, this fact may be stated mathematically by saying that the probability of being in a particular state on the k^{th} observation does depend on some or all of the previous $k-1$ states which were observed.

Example 9.1.2. Suppose we draw without replacement 13 cards from an ordinary bridge deck. Then the probability that the eighth draw produces an ace of spades depends on the outcome of the preceding 7 draws. For this system there are 52 possible states. let $E_k = \{$ ace of spades $\}$. Then

$$\mathrm{P}\left(X_8 = E_k | X_1 = E_{j1}, \ldots, X_7 = E_{j7}\right) = \begin{cases} \frac{1}{45} & X_i \neq E_k \ \forall i = 1, 2, \ldots, 7 \\ 0 & \text{otherwise} \end{cases}. \qquad \square$$

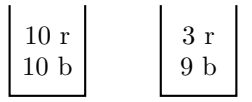
Definition 9.1.1 (Markov Chain). A *Markov Chain* is a probabilistic model that applies to systems that exhibit a special type of dependence, where the state of the system on the k^{th} observation depends only on the state of the system on the $(k-1)^{\text{st}}$ observation. That is,

$$\mathrm{P}\left(X_k = E_i | X_1 = E_{j1}, \ldots, X_{k-1} = E_{jk-1}\right) = \mathrm{P}\left(X_k = E_i | X_{k-1} = E_{jk-1}\right) \quad \forall k \in \mathbb{Z}^*.$$

Once this type of system is in a given state, future changes in the system depend only on this state and not on the manner in which the system arrived at this particular state.

Example 9.1.3. Consider two urns: One red and one black:

The red urn contains 10 red and 10 black balls,
The black urn contains 3 red and 9 black balls:



red urn black urn

The system begins with the red urn where a ball is drawn, its color is noted, and it is then replaced. If the ball drawn is red, the second draw is from the red urn; if the ball drawn is black, the second draw is from the black urn. This process is repeated with the urn chosen for a draw determined by the color of the ball on the previous draw. The two possible states in this example are $E_0 = \{$ red ball $\}$ and $E_1 = \{$ black ball $\}$. Assume that when drawing from an urn, each ball in that urn has the same probability of being drawn. The probability that on the 5^{th} drawing we obtain a red ball given the information that the outcomes of the previous drawings were (black, black, red, black) $= (E_1, E_1, E_0, E_1)$ is simply the probability of a red ball on the fifth draw, given that the fourth draw produced a black ball. That is,

$$\mathrm{P}(X_5 = E_0 | X_1 = E_1, X_2 = E_1, X_3 = E_0, X_4 = E_1) = \mathrm{P}(X_5 = E_0 | X_4 = E_1) = \frac{3}{12} = \frac{1}{4}.$$

Note that the above result contrasts with

$$\mathrm{P}(X_5 = E_0 | X_1 = E_1, X_2 = E_1, X_3 = E_0, X_4 = E_0) = \mathrm{P}(X_5 = E_0 | X_4 = E_0) = \frac{10}{20} = \frac{1}{2}. \qquad \square$$

9.2 Basic Concepts of Markov Chain Theory

Definition 9.2.1. The *state space* S of a Markov Chain is the set of all possible (and perhaps some impossible) states of the system. I.e., it is the sample space.

For Example 9.1.3, the state space $S = \{\{\text{red ball}\}, \{\text{black ball}\} = \{E_1, E_2\}$.

The states of a Markov chain are exclusive of one another – i.e., no two states can occur at the same time. markov chains are applicable only to systems where the number of states is finite or countably infinite.

Definition 9.2.2. A Markov chain is *finite* if the state space of the chain is finite.

When making observations of a system whose probabilistic model is a Markov chain, instead of saying that the i^{th} observation resulted in E_k, we usually say that at time i the system was in state E_k, or more simply in state k. There are various conventions for starting time. Some writers start with time 1 so that a sample is denoted by $(X_1, X_2, \ldots, X_n)$. Others starts with time 0 so that a sample is denoted $(X_0, X_1, \ldots, X_n)$. Here we shall start with time 0.

Definition 9.2.3. The *one-step transition probability function* (homogeneous)[1] for a Markov chain is a function that gives the probability of going from state j to state k in one step (one time interval) for each j and k. The one-step transition probability function in the homogeneous case is given by

$$p_{jk} = \mathrm{P}(X_n = E_k | X_{n-1} = E_j) \quad \forall j, k \ni E_j, E_k \in S, \quad \forall n \geq 2.$$

Many authors denote E_j and E_k by j and k, respectively:

$$p_{jk} = \mathrm{P}(E_k | E_j) = \mathrm{P}(k|j).$$

The one-step transition probabilities can be arranged into a matrix form, called the *transition probability matric* (tpm), as follows:

$$\mathbf{P} = \begin{array}{c} \nearrow \\ E_0 \\ E_1 \\ \vdots \\ E_i \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} E_0 & E_1 & \cdots & E_j & \cdots \end{array} \\ \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0j} & \cdots \\ p_{10} & p_{11} & \cdots & p_{1j} & \cdots \\ \vdots & \vdots & & \vdots & \\ p_{i0} & p_{i1} & \cdots & p_{ij} & \cdots \\ \vdots & \vdots & & \vdots & \end{bmatrix} \end{array}.$$

In the above notation, we label the states on the left and upper borders of the matrix, plus add "$\nearrow$" to emphasize that we are going from the left side to above.

Example 9.2.1. In Example 9.1.1,

$$\begin{array}{l} p_{00} = \mathrm{P}(E_0|E_0) = \frac{1}{2} \\ p_{01} = \mathrm{P}(E_1|E_0) = \frac{1}{2} \\ p_{10} = \mathrm{P}(E_0|E_1) = \frac{1}{4} \\ p_{11} = \mathrm{P}(E_1|E_1) = \frac{3}{4} \end{array} \quad \Rightarrow \quad \mathbf{P} = \begin{array}{c} \nearrow \\ E_0 \\ E_1 \end{array} \begin{array}{c} \begin{array}{cc} E_0 & E_1 \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} \end{array}.$$

We see that the individual rows of $\mathbf{P}$ sum to 1. That is, $\frac{1}{2} + \frac{1}{2} = 1$ and $\frac{1}{4} + \frac{3}{4} = 1$. This is not a peculiarity of this particular example, but is true of all one-step tpms $\mathbf{P}$. □

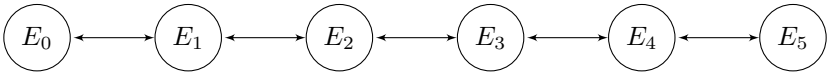
Definition 9.2.4. A square matrix $\mathbf{P}$ is a *stochastic matrix* if for every row i,

$$\sum_{\text{all } j} p_{ij} = 1.$$

All tpms are stochastic matrices.

[1] In general, the one-step transition probability function is given by $p_{jk}(n-1, n) = \mathrm{P}(X_n = E_k | X_{n-1} = E_j)$.)

Consider a particle that moves in a straight line in unit steps with the probability of a step to the right given by p and a step to the left given by $q = 1 - p$. This type of system is called a *random walk*, and variations on it afford some interesting examples of finite Markov chains. Let the number of possible states be 6. The state space for this chain is illustrated below. the variations on the random walk model will be for the boundary (barrier) states, which are states 0 and 5.



Example 9.2.2 (Random Walk with Reflecting Barriers). If the particle is in state 0, it moves to the right with probability p and stays there with probability q. If the particle is in state 5, it moves to the left with probability q and stays there with probability p. The tpm of this is

$$\mathbf{P} = \begin{array}{c} \nearrow \\ E_0 \\ E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \end{array} \begin{array}{c} \begin{array}{cccccc} E_0 & E_1 & E_2 & E_3 & E_4 & E_5 \end{array} \\ \begin{bmatrix} q & p & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 & 0 \\ 0 & q & 0 & p & 0 & 0 \\ 0 & 0 & q & 0 & p & 0 \\ 0 & 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & q & p \end{bmatrix} \end{array}. \qquad \square$$

Example 9.2.3 (Random Walk with Absorbing Barriers). If the particle is in state 0 or 5, it stays in that state. Thus, we set $p_{00} = p_{55} + 1$:

$$\mathbf{P} = \begin{array}{c} \nearrow \\ E_0 \\ E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \end{array} \begin{array}{c} \begin{array}{cccccc} E_0 & E_1 & E_2 & E_3 & E_4 & E_5 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 & 0 \\ 0 & q & 0 & p & 0 & 0 \\ 0 & 0 & q & 0 & p & 0 \\ 0 & 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}. \qquad \square$$

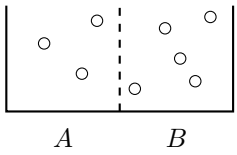
Example 9.2.4. For a specific example of a random walk with absorbing barriers, consider the game of matching pennies. Assume that two opponents, Nate and D.J., have five pennies between them. The coins are fair so that the probabilities of a head or a tail are equal (i.e., equal to $\frac{1}{2}$). Nate tosses a coin and records the outcome. Then D.J. tosses a coin. If Nate's and D.J.'s coins are the same (i.e., both heads or both tails), then Nate wins D.J.'s coin; otherwise, D.J. wins it. Thus, on each toss, Nate wins with probability $\frac{1}{2}$. Let $E_i = \{i\}$ represent the number of pennies Nate has won. Thus, our states are $\{E_i : \ i = 0, 1, 2, 3, 4, 5\}$. The game ends with Nate has either all (5) or none (0) of the pennies. This is a random walk with $p = q = \frac{1}{2}$. Given that Nate is in state k (he has k pennies), he either goes to state $k + 1$ (wins) or $k - 1$ (loses), each with probability $\frac{1}{2}$. States 0 and 5 are absorbing states:

$$\mathbf{P} = \begin{array}{c} \nearrow \\ E_0 \\ E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \end{array} \begin{array}{c} \begin{array}{cccccc} E_0 & E_1 & E_2 & E_3 & E_4 & E_5 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}.$$

In the above matrix, $p_{34} = \frac{1}{2}$ is the probability that Nate wins a fourth coin, given that he currently has 3 coins. □

Example 9.2.5 (Ehrenfest Diffusion Model). The *Ehrenfest Diffusion Model* is another example of a random walk problem. Consider a physical system where k molecules are distributed between two containers A and B. These containers are separated by a permeable membrane so that the molecules can move freely between them:

At each time point t, one of k molecules is chosen at random and moved from one container to the other.



Mathematically, we set up the model by assuming that we have containers A and B, where movement is achieved in the following manner: At each instant of time t, one of the k molecules is chosen at random and moved from one container to the other. the state of the system is determined by the number of molecules in A and in B. Let the state $E_i = \{i$ molecules are in $A\}$. The possible states of the system are then $E_0, \ldots, E_k$. In the Ehrenfest model, if A has j molecules, it can on the next trial go to states E_{j+1} or E_{j-1} with probabilities $\frac{j}{k}$ and $\frac{k-j}{k}$, respectively:

$$\mathrm{P}(E_{j-1}|E_j) = p_{j,j-1} = \tfrac{j}{k}, \quad \mathrm{P}(E_{j+1}|E_j) = p_{j,j+1} = \tfrac{k-j}{k}.$$

The tpm is thus given by

$$\mathbf{P} = \begin{array}{c} \nearrow \\ E_0 \\ E_1 \\ E_2 \\ \vdots \\ E_{k-1} \\ E_k \end{array} \begin{array}{c} \begin{array}{ccccccc} E_0 & E_1 & E_2 & E_3 & \cdots & E_{k-1} & E_k \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ \frac{1}{k} & 0 & \frac{k-1}{k} & 0 & \cdots & 0 & 0 \\ 0 & \frac{2}{k} & 0 & \frac{k-2}{k} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \frac{1}{k} \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \end{array}.$$

These transition probabilities show a tendency to shift toward an equilibrium of 50% of the molecules in each onctainer. □

Definition 9.2.5. The *initial probability function* is a function that gives the probability that the system is initially (at tome zero) in state i, $\forall i$. The initial probability function will be denoted by

$$p_i^{(0)} = \mathrm{P}(X_0 = E_i) \quad \forall i.$$

The initial probability function is commonly arrayed in a vector form:

$$\mathbf{p}^{(0)} = (p_0^{(0)}, p_1^{(0)}, \ldots).$$

Example 9.2.6. In Example 9.1.3, the initial condition was that drawing was to begin in the red urn. Thus,

$$\begin{array}{l} p_0^{(0)} = \mathrm{P}(X_0 = E_0) = \frac{1}{2}, \\ p_1^{(0)} = \mathrm{P}(X_0 = E_1) = \frac{1}{2}. \end{array} \quad \Rightarrow \quad \mathbf{p}^{(0)} = \left(\tfrac{1}{2}, \tfrac{1}{2}\right). \qquad \square$$

9.3 Describing a System by a Markov Chain

A Markov chain is completely described when the state space S, initial probability function $\mathbf{p}^{(0)}$, and one-step transition probability function $\mathbf{P}$ are given. Therefore, to represent a physical system by a Markov chain, each one of these must be calculated or estimated. Once this is done, there are four principal questions that we can investigate using this Markov chain:

1. What is the probability of going from state j to state k in n steps?
2. What is the unconditional probability that at time n (i.e., n steps after the last observation), the system is in state j?
3. If a Markov chain terminates when it reaches a state k (which by definition is an absorbing state), then what is the expected (mean) time to reach k (and thus terminate the chain), given that the chain started in state j?
4. Is there a steady" state behavior for a Markov chain? That is, for each state j, does there exist a probability function $\pi_j = \lim_{n\to\infty} \mathrm{P}(X_n = E_j)$

Each of these four questions can be answered using the information contained in S, $\mathbf{p}^{(0)}$ and $\mathbf{P}$.

9.3.1 *n*-step Transition Probability Matrix

We want to derive an n-step transition probability function

$$p_{jk}^{(n)} = \mathrm{P}(X_{t+n} = E_k | X_t = E_j).$$

In n steps, a system may go from E_j to E_k by a number of different paths. For example, if the system has r possible states, then in two steps it may go from E_j to E_k by

$$\begin{array}{c} E_j \to E_0 \to E_k, \\ E_j \to E_1 \to E_k, \\ \vdots \\ E_j \to E_{r-1} \to E_k. \end{array} \tag{9.3.1}$$

In order to compute the probability of the event $E_j \to E_i \to E_k$, we need independence. As mentioned in chapter XX, two events A and B are independent if $\mathrm{P}(AB) = \mathrm{P}(A)\mathrm{P}(B)$.

Example 9.3.1. Toss a coin twice. For H = head and T = tail, suppose $\mathrm{P}(H) = \mathrm{P}(T) = \frac{1}{2}$. Let X_i denote the outcome of the i^{th} toss. We know by assumption that X_1 and X_2 are independent. Therefore,

$$\mathrm{P}(X_1 = H, X_2 = H) = \mathrm{P}(X_1 = H)\mathrm{P}(X_2 = H) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}. \qquad \square$$

For Markov chains, the event of going from E_j to E_i in one step, $E_j \to E_i$, is independent of the event $E_i \to E_k$. This follows from the definition of a Markov chain. Thus,

$$\mathrm{P}([E_j \to E_i] \cap [E_i \to E_k]) = \mathrm{P}(E_j \to E_i)\mathrm{P}(E_i \to E_k) = p_{ji}p_{ik}.$$

Since $\mathrm{P}(E_j \to E_i \to E_k) = \mathrm{P}([E_j \to E_i] \cap [E_i \to E_k])$, we now have expressions for computing the probabilities of the r paths listed in (9.3.1). Since these r paths are mutually exclusive (i.e., no pair of them happen simultaneously), $p_{ji}^{(2)}$ is equal to the sum of the probabilities over these r different paths:

$$p_{jk}^{(2)} = \sum_{i \in S} p_{ji}p_{ik}.$$

By the same logic, the probability $p_{jk}^{(3)}$ of going from j to k in three steps is

$$p_{jk}^{(3)} = \sum_{i \in S} p_{ji}p_{ik}^{(2)}.$$

By induction on n, it follows that

$$p_{jk}^{(n)} = \sum_{i \in S} p_{ji}p_{ik}^{(n-1)}. \tag{9.3.2}$$

We now have an answer to question 1 of this section: What is the probability of going from state j to state k in n steps?

Using the same reasoning as above, we can express $p_{jk}^{(n)}$ in a slightly different form:

$$p_{jk}^{(n)} = \sum_{i \in S} p_{ji}^{(n-1)}p_{ik}. \tag{9.3.3}$$

More generally, using induction we can show that

$$p_{jk}^{(m+n)} = \sum_{i \in S} p_{ji}^{(m)}p_{ik}^{(n)}. \tag{9.3.4}$$

This is a special case of the *Chapman-Kolmogorov equation*. We can expressed these n-step transition probabilities in a matrix form:

$$\mathbf{P}^{(n)} = \begin{array}{c} \nearrow \\ E_0 \\ E_1 \\ \vdots \\ E_i \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} E_0 & & E_1 & \cdots & E_j & \cdots \end{array} \\ \begin{bmatrix} p_{00}^{(n)} & p_{01}^{(n)} & \cdots & p_{0j}^{(n)} & \cdots \\ p_{10}^{(n)} & p_{11}^{(n)} & \cdots & p_{1j}^{(n)} & \cdots \\ \vdots & \vdots & & \vdots & \\ p_{i0}^{(n)} & p_{i1}^{(n)} & \cdots & p_{ij}^{(n)} & \cdots \\ \vdots & \vdots & & \vdots & \end{bmatrix} \end{array}.$$

Algebraically, we can show (not here) that $\mathbf{P}^{(n)} = \mathbf{P}^n$, where $\mathbf{P}^n$ is the one-step tpm multiplied by itself n times. In matrix form, we can write (9.3.2) as

$$\mathbf{P}^n = \mathbf{P}\mathbf{P}^{n-1},$$

(9.3.3) as

$$\mathbf{P}^n = \mathbf{P}^{n-1}\mathbf{P},$$

and (9.3.4) as

$$\mathbf{P}^{m+n} = \mathbf{P}^m\mathbf{P}^{n-1}.$$

Example 9.3.2. From Example 9.2.1, we have

$$\mathbf{P} = \begin{array}{c} \nearrow \\ E_0 \\ E_1 \end{array} \begin{array}{c} \begin{array}{cc} E_0 & E_1 \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} \end{array}.$$

Therefore,

$$\mathbf{P}^2 = \mathbf{P}\cdot\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} = \begin{bmatrix} \frac{1}{4}+\frac{1}{8} & \frac{1}{4}+\frac{3}{8} \\ \frac{1}{8}+\frac{3}{16} & \frac{1}{8}+\frac{9}{16} \end{bmatrix} = \begin{bmatrix} \frac{3}{8} & \frac{5}{8} \\ \frac{5}{16} & \frac{11}{16} \end{bmatrix}.$$

and

$$\mathbf{P}^3 = \mathbf{P}\cdot\mathbf{P}^2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}\begin{bmatrix} \frac{3}{8} & \frac{5}{8} \\ \frac{5}{16} & \frac{11}{16} \end{bmatrix} = \begin{bmatrix} \frac{3}{16}+\frac{5}{32} & \frac{5}{16}+\frac{11}{32} \\ \frac{3}{32}+\frac{15}{64} & \frac{5}{32}+\frac{33}{64} \end{bmatrix} = \begin{bmatrix} \frac{11}{32} & \frac{21}{32} \\ \frac{21}{64} & \frac{43}{64} \end{bmatrix}. \qquad \square$$

9.3.2 Unconditional Probability Functions

To get the unconditional probability function $p_k^{(n)} = \mathrm{P}(X_n = E_k)$ of being in state k in n steps, we can use a slightly different form of (9.3.4):[2]

$$p_k^{(n)} = \sum_{i\in S} p_i^{(0)} p_{ik}^{(n)}. \tag{9.3.5}$$

These unconditional probabilities $p_k^{(n)}$ can be written in vector form as

$$\mathbf{p}^{(n)} = (p_0^{(n)}, p_1^{(n)}, \ldots).$$

Therefore, we can write (9.3.5) in matrix-vector form as

$$\mathbf{p}^{(n)} = \mathbf{p}^{(0)}\mathbf{P}^n.$$

Recall that $\mathbf{P}$, $\mathbf{P}^{(n)}$ and $\mathbf{P}^n$ refer to (*conditional*) transition probability *matrices*, while $\mathbf{p}$ and $\mathbf{p}^{(n)}$ refers to an *unconditional* probability *vectors*. There also should be no confusion between the *conditional* probability $p_{jk}^{(n)}$ of going from states j to k in n steps, and the *unconditional* probability $p_k^{(n)}$ of being at state k in n steps.

[2] which we can derive using the same logic as for (9.3.4).

Example 9.3.3. As a continuation of Example 9.3.2, now assume that in addition to $\mathbf{P}$, we also have an initial probability function $\mathbf{p}^{(0)}$:

$$\mathbf{P} = \begin{array}{c} \nearrow \\ E_0 \\ E_1 \end{array} \begin{array}{c} \begin{array}{cc} E_0 & E_1 \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} \end{array}, \qquad \mathbf{p}^{(0)} = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \end{bmatrix}.$$

Therefore,

$$\mathbf{p}^{(1)} = \mathbf{p}^{(0)}\mathbf{P} = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} = \begin{bmatrix} \left(\frac{3}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{4}\right) & \left(\frac{3}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{3}{4}\right) \end{bmatrix} = \begin{bmatrix} \frac{7}{16} & \frac{9}{16} \end{bmatrix}.$$

and

$$\mathbf{p}^{(2)} = \mathbf{p}^{(0)}\mathbf{P}^2 = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} \frac{3}{8} & \frac{5}{8} \\ \frac{5}{16} & \frac{11}{16} \end{bmatrix} = \begin{bmatrix} \left(\frac{3}{4} \cdot \frac{3}{8} + \frac{1}{4} \cdot \frac{5}{16}\right) & \left(\frac{3}{4} \cdot \frac{5}{8} + \frac{1}{4} \cdot \frac{11}{16}\right) \end{bmatrix} = \begin{bmatrix} \frac{23}{64} & \frac{41}{64} \end{bmatrix}. \qquad \square$$

We now have the answer to question 2 of this section: What is the unconditional probability that at time n (i.e., n steps after the last observation), the system is in state j?

9.3.3 Classification of States

Definition 9.3.1. A state k is *accessible* from a state i if $\exists$ a positive integer $n \ni p_{jk}^{(n)} > 0$. In other words, it is possible to go from state i to state j in some (finite) number of steps.

Example 9.3.4. Suppose we have the following tpm:

$$\mathbf{P} = \begin{array}{c} \nearrow \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} q & p & 0 & 0 \\ q & 0 & p & 0 \\ 0 & q & 0 & p \\ 0 & 0 & q & p \end{bmatrix} \end{array}.$$

State 3 is accessible from state 2, since $p_{23}^{(1)} = p_{23} > 0$. In fact, every state is accessible from every state (including itself). $\square$

Definition 9.3.2. Two states j and k *communicate* if j is accessible from k and k is accessible from j.

Example 9.3.5. In Example 9.3.4, states 2 and 3 communicate, since $p_{23} > 0$ and $p_{32} > 0$. In fact, every state communicates with every state. $\square$

Definition 9.3.3. A nonempty set C of states is *closed* if no state outside the set is accessible from any state inside the set. A single state k forming a closed set is an *absorbing state*. A Markov chain that has one or more absorbing states is said to be an *absorbing Markov chain*.

Once a Markov chain enters a closed set, it remains within that set.

Definition 9.3.4. A Markov chain is *irreducible* if all pairs of states communicate.

If a Markov chain has a closed set C, then for any state $j \in C$ and $k \notin C$, $p_{jk}^{(n)} = 0 \ \forall n$.

Example 9.3.6. Suppose we have the tpm

$$\mathbf{P} = \begin{array}{c} \nearrow \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{array} \begin{array}{c} \begin{array}{cccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{8} & \frac{1}{8} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 \\ \frac{1}{6} & 0 & 0 & 0 & 0 & \frac{2}{3} & 0 & 0 & \frac{1}{6} & 0 \\ \frac{2}{3} & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{6} \end{bmatrix} \end{array}.$$

Closed sets are $C_1 = \{0\}$, $C_2 = \{1, 2\}$, and $C_3 = \{3, 4, 5, 6\}$. In fact, the three submatrices whose states form closed sets may be studied separately:

$$\begin{array}{c} \nearrow \\ 0 \end{array} \begin{array}{c} \begin{array}{c} 0 \end{array} \\ \begin{bmatrix} 1 \end{bmatrix} \end{array}, \qquad \begin{array}{c} \nearrow \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{array}, \qquad \begin{array}{c} \nearrow \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccc} 3 & 4 & 5 & 6 \end{array} \\ \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ 1 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{8} & \frac{1}{8} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \end{array}.$$

In particular,

$$\mathbf{P}^n = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \multicolumn{2}{c}{\multirow{2}{*}{$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}^n$}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \multicolumn{4}{c}{\multirow{4}{*}{$\begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ 1 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{8} & \frac{1}{8} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}^n$}} & 0 & 0 & 0 \\ 0 & 0 & 0 & & & & & 0 & 0 & 0 \\ 0 & 0 & 0 & & & & & 0 & 0 & 0 \\ 0 & 0 & 0 & & & & & 0 & 0 & 0 \\ * & * & * & * & * & * & * & * & * & * \end{bmatrix}$$

This Markov chain is not irreducible. □

Definition 9.3.5. A state j is *transient*, or *nonrecurrent* if the conditional probability of returning to j given that the system starts at j is less than one.

It can be shown (not here) that a state j being transient is equivalent to

$$\sum_{n=1}^{\infty} p_{jj}^{(n)} < \infty \quad \text{or} \quad \sum_{n=1}^{\infty} p_{kj}^{(n)} < \infty \ \forall k \in S.$$

It can be shown that if an irreducible Markov chain has a transient state, it must possess infinitely many states; i.e., if a chain has a transient state, it is not irreducible.

Definition 9.3.6. The *period* of a state k in a finite Markov chain is the greatest common divisor of the set of positive integers n for which $p_{kk}^{(n)} > 0$. A state k is called *aperiodic* if it has period 1.

Example 9.3.7. Let

$$\mathbf{P} = \begin{array}{c} \nearrow \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{array}$$

represent the tpm of a two-state Markov chain. Then

$$\mathbf{P}^2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{P}^4 = \mathbf{P}^{2m} \quad \forall m \in \mathbb{Z}^+.$$

Thus, $p_{11}^{(n)} > 0 \ \forall n = 2k, \ k \in \mathbb{Z}^+$. The greatest common divisor of this set is 2. Hence state 0 has period 2, as does state 1. This is an irreducible periodic Markov chain. □

Definition 9.3.7. A finite Markov chain is called *aperiodic* if there exists a state k with period 1.

We can show that a Markov chain is aperiodic by exhibiting a state k for which $p_{kk} = p_{kk}^{(1)} > 0$. Thus if any of the diagonal elements in $\mathbf{P}$ are non-zero, the chain is aperiodic. The converse is not true, however, as a chain may be aperiodic even if all the diagonal elements of $\mathbf{P}$ are zero. We can also show (not here) that a Markov chain is aperiodic by showing that there exists an integer n such that $p_{jk}^{(n)} > 0 \ \forall j, k$.

Example 9.3.8. If

$$\mathbf{P} = \begin{array}{c} \nearrow \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 \end{bmatrix} \end{array},$$

then the chain is aperiodic, since $p_{11} = \frac{1}{2} > 0$. If

$$\mathbf{P} = \begin{array}{c} \nearrow \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \end{array},$$

the chain is still aperiodic, since

$$\mathbf{P}^2 = \begin{array}{c} \nearrow \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix} \end{array}.$$

and thus $p_{jk}^{(2)} > 0 \ \ \forall j, k \in S$.

Intuitively, we can perhaps understand why this second tpm $\mathbf{P}$ is aperiodic: Pick any state – say, 2. Since $p_{22}^{(2)} > 0$, we know we can go from state 2 to state 2 in $n = 2$ steps. However, since $p_{21}^{(2)} > 0$ and $p_{12}^{(1)} > 0$, we know that we can go from state 2 to state 1 in two steps, and then to state 2 in one step – thus going from state 2 to state 2 in $n = 3$ steps. Altogether, we can go from 2 to 2 in $n = 2$ or $n = 3$ steps. The gcd of $\{2, 3\}$ is 1.

Indeed, this logic can be used for any state in $S = \{0, 1, 2\}$. □

9.3.4 Absorption Times for Finite Markov Chains

Let $\mathbf{P}$ denote the tpm for a finite absorbing Markov chain consisting of a set of transient states T and the set of absorbing states. $\mathbf{P}$ may be written as

$$\mathbf{P} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{R} & \mathbf{Q} \end{bmatrix},$$

where $\mathbf{I}$ is the identity transition matrix representing the absorbing states, $\mathbf{Q} = \{p_{jk}\}$ is a square matrix where j and k are members of the transient set T, and $\mathbf{R}$ is a rectangular matrix of transition probabilities from the transient states to the absorbing states. Since the Markov chain is finite, T is not a closed communicating set (irreducible) and hence does not determine a sub-Markov chain. Thus, $\mathbf{Q}$ is not a stochastic sub-matrix.

Example 9.3.9. Let

$$\mathbf{P} = \begin{array}{c} \nearrow \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 \\ 0 & q & 0 & p & 0 \\ 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \xrightarrow{p=\frac{1}{4}} \mathbf{P} = \begin{array}{c} \nearrow \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{3}{4} & 0 & \frac{1}{4} & 0 & 0 \\ 0 & \frac{3}{4} & 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{3}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}.$$

Using the state transformation $0 \to 0'$, $4 \to 1'$, $1 \to 2'$, $2 \to 3'$ and $3 \to 4'$, we rewrite $\mathbf{P}$ as

$$\mathbf{P} = \begin{array}{c} \nearrow \\ 0' \\ 1' \\ 2' \\ 3' \\ 4' \end{array} \begin{array}{c} \begin{array}{ccccc} 0' & 1' & 2' & 3' & 4' \end{array} \\ \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hdashline \frac{3}{4} & 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{3}{4} & 0 & \frac{1}{4} \\ 0 & \frac{1}{4} & 0 & \frac{3}{4} & 0 \end{array}\right] \end{array} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{R} & \mathbf{Q} \end{bmatrix}$$

with

$$\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{0} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} \frac{3}{4} & 0 \\ 0 & 0 \\ 0 & \frac{1}{4} \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 0 & \frac{1}{4} & 0 \\ \frac{3}{4} & 0 & \frac{1}{4} \\ 0 & \frac{3}{4} & 0 \end{bmatrix}.$$

□

Let $\mathbf{p}^{(0)}$ be the vector of initial probabilities. This vector can be partitioned as

$$\mathbf{p}^{(0)} = \begin{bmatrix} \mathbf{p}_A^{(0)} & \mathbf{p}_t^{(0)} \end{bmatrix},$$

where $\mathbf{p}_A^{(0)}$ and $\mathbf{p}_T^{(0)}$ are the vectors that give the probabilities of being initially in each of the absorbing or transient states, respectively. For most physical systems studied, $\mathbf{p}_A^{(0)} = \mathbf{0}$, the zero vector. That is, the probabilitiy of initially being in an absorbing state is zero.

Recall that the physical system represented by a Markov chain is observed periodically at times $0, 1, 2, 3 \ldots$. Let the variable W represent the time to absorption (that is, the number of steps to absorption) from the set of transient states to the set of absorbing states. This variable can take values $1, 2, 3, 4, \ldots$.

The probability of going from the set of transient states to an absorbing state in n steps can be shown (not here) to be

$$\mathrm{P}(W = n) = \mathbf{p}_T^{(0)} \mathbf{Q}^{n-1} \mathbf{R1}, \quad n \in \mathbb{Z}^+, \tag{9.3.6}$$

where $\mathbf{1}$ is a column vector whose components are all equal to 1. Let r and s denote the number of total and transient states, respectively. Then $\mathbf{p}_T^{(0)}$ is a $1 \times s$ matrix and $\mathbf{Q}^{n-1}$ is an $s \times s$ matrix that gives the probability of staying in the transient states for $n-1$ time intervals. $\mathbf{R}$ is an $s \times (r-s)$ matrix giving the probability of going from a transient state to an absorbing state. Multiplying $\mathbf{R}$ by the column vector $\mathbf{1}$ simply sums each row of $\mathbf{R}$ so that $\mathbf{R1}$ is an $s \times 1$ column vector that gives for each transient state j the probability of going from j to the set of aobsorbing states. The matrix multiplication $\mathbf{p}_T^{(0)} \mathbf{Q}^{n-1} \mathbf{R1}$ simply sums up, for each transient state $j \in T$, the probability of starting in j, staying in the set of transient states for $n-1$ steps, and then going into the set of absorbing states.

Example 9.3.10. Using the transformed transition matrix $\mathbf{P}$ of Example 9.3.9, $\mathbf{R}$ and $\mathbf{Q}$ are given as

$$\mathbf{R} = \begin{bmatrix} \frac{3}{4} & 0 \\ 0 & 0 \\ 0 & \frac{1}{4} \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 0 & \frac{1}{4} & 0 \\ \frac{3}{4} & 0 & \frac{1}{4} \\ 0 & \frac{3}{4} & 0 \end{bmatrix}.$$

Let $\mathbf{p}^{(0)} = \begin{bmatrix} 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$, Then $\mathbf{p}_T^{(0)} = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$. By equation (9.3.6), the probability of going from the set of transient states to an absorbing state in exactly one step is

$$\mathrm{P}(W = 1) = \mathbf{p}_T^{(0)} \mathbf{Q}^0 \mathbf{R1} = \mathbf{p}_T^{(0)} \mathbf{R1} = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} \frac{3}{4} & 0 \\ 0 & 0 \\ 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} \frac{3}{4} \\ 0 \\ \frac{1}{4} \end{bmatrix} = \frac{3}{16} + \frac{1}{16} = \frac{1}{4}.$$

The probability of going from the set of transient states to an absorbing state in exactly two steps is

$$\mathrm{P}(W = 2) = \mathbf{p}_T^{(0)} \mathbf{Q}^1 \mathbf{R1} = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{4} & 0 \\ \frac{3}{4} & 0 & \frac{1}{4} \\ 0 & \frac{3}{4} & 0 \end{bmatrix} \begin{bmatrix} \frac{3}{4} \\ 0 \\ \frac{1}{4} \end{bmatrix} = \frac{5}{16};$$

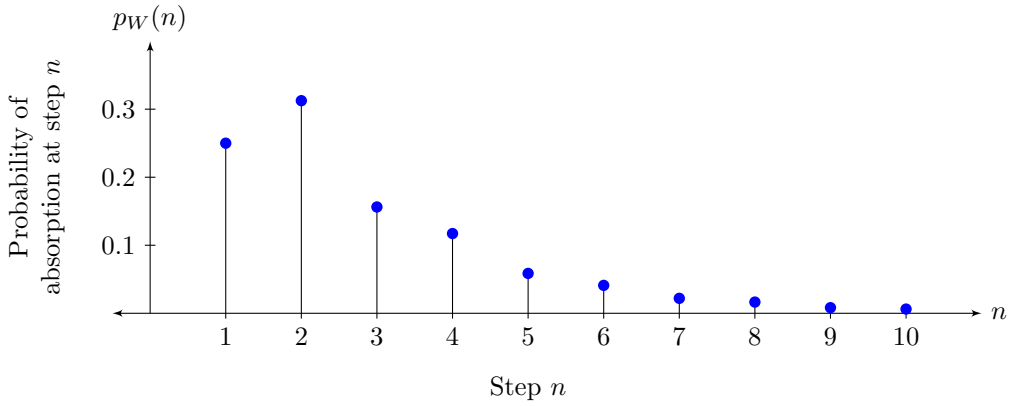
The probability of going from the set of transient states to an absorbing state in exactly three steps is

$$\mathrm{P}(W = 3) = \mathbf{p}_T^{(0)} \mathbf{Q}^2 \mathbf{R1} = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{4} & 0 \\ \frac{3}{4} & 0 & \frac{1}{4} \\ 0 & \frac{3}{4} & 0 \end{bmatrix}^2 \begin{bmatrix} \frac{3}{4} \\ 0 \\ \frac{1}{4} \end{bmatrix} = \frac{5}{32};$$

The probability of going from the set of transient states to an absorbing state in exactly four steps is

$$\mathrm{P}(W=4) = \mathbf{p}_T^{(0)}\mathbf{Q}^3\mathbf{R}\mathbf{1} = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{4} & 0 \\ \frac{3}{4} & 0 & \frac{1}{4} \\ 0 & \frac{3}{4} & 0 \end{bmatrix}^3 \begin{bmatrix} \frac{3}{4} \\ 0 \\ \frac{1}{4} \end{bmatrix} = \frac{15}{128}.$$

Continuing this process, we can graph the values of $p_W(n) = \mathrm{P}(W=n),\ n \in \mathbb{Z}^+$, which is a pmf:



□

In Example 9.3.10, we asserted without proof that $p_W(n) = \mathrm{P}(W=n)$ is indeed a pmf. This is proven easily enough – since all entries of the component matrices are nonnegative, we know that $\mathrm{P}(W=n) \geq 0\ \forall n$. Furthermore, since the events $\{W=i\}$ and $\{W=j\}$ for $i \neq j$ cannot occur simultaneously, and there must be a transition from the transient states to the absorbing states in some number n of steps, it follows that

$$\sum_{n=0}^{\infty} p_W(n) = \sum_{n=0}^{\infty} \mathrm{P}(W=n) = 1.$$

In the graph above, we can see that $\lim_{n\to\infty} \mathrm{P}(W=n) = 0$, which is true of any pmf.

Geometrically, this means that if the line segments connecting the points $(n,0)$ and $(n,\mathrm{P}(W=n))$ are summed up $\forall n \in \mathbb{Z}^+$, the sum would have unit length.

The mean time (expected time) of absorption from the set of transient states is defined mathematically as

$$\mu_W = \mathrm{E}[W] = \sum_{n=1}^{\infty} n\mathrm{P}(W=n).$$

It can be shown (not here) that the mean time of absorption from the set of transient states is given by

$$\mu_W = \mathrm{E}[W] = \mathbf{p}_T^{(0)}(\mathbf{I}-\mathbf{Q})^{-1}\mathbf{1}. \tag{9.3.7}$$

It is common to let

$$\mathbf{N} = (\mathbf{I}-\mathbf{Q})^{-1}. \tag{9.3.8}$$

This matrix is called the *fundamental matrix of an absorbing Markov chain.* It can be verified that $\mathbf{I}-\mathbf{Q}$ has an inverse by showing that

$$(\mathbf{I}-\mathbf{Q})^{-1} = \sum_{k=0}^{\infty} \mathbf{Q}^k$$

and that this infinite series converges.

The second moment of absorption time is defined to be $\sum_{n=1}^{\infty} n^2 \mathrm{P}(W=n)$ and can be shown (not here) to be given by

$$\mathrm{E}\left[W^2\right] = \mathbf{p}_T^{(0)}(2\mathbf{N}-\mathbf{I})\mathbf{N}\mathbf{1}. \tag{9.3.9}$$

Mathematically, the variance for a pdf of absorption time is given by

$$\mathrm{Var}(W) = \sum_{n=1}^{\infty}(n-\mu_W)^2\mathrm{P}(W=n) = \mathrm{E}\left[W^2\right] - (\mathrm{E}[W])^2 = \mathbf{p}_T^{(0)}(2\mathbf{N}-\mathbf{I})\mathbf{N}\mathbf{1} - \left(\mathbf{p}_T^{(0)}\mathbf{N}\mathbf{1}\right)^2. \tag{9.3.10}$$

Example 9.3.11. As a continuation of Example 9.3.10, we will compute the mean and variance of the absorption time, using (9.3.7) and (9.3.10). First we compute

$$\mathbf{I}-\mathbf{Q} = \begin{bmatrix}1&0&0\\0&1&9\\0&0&1\end{bmatrix} - \begin{bmatrix}0&\frac{1}{4}&0\\\frac{3}{4}&0&\frac{1}{4}\\0&\frac{3}{4}&0\end{bmatrix} = \begin{bmatrix}1&-\frac{1}{4}&0\\-\frac{3}{4}&1&-\frac{1}{4}\\0&-\frac{3}{4}&1\end{bmatrix}$$

so that

$$\mathbf{N} = (\mathbf{I}-\mathbf{Q})^{-1} = \begin{bmatrix}\frac{13}{10}&\frac{2}{5}&\frac{1}{10}\\\frac{6}{5}&\frac{8}{5}&\frac{2}{5}\\\frac{9}{10}&\frac{6}{5}&\frac{13}{10}\end{bmatrix}.$$

Therefore, the mean absorption time is

$$\mathrm{E}[W] = \mathbf{p}_T^{(0)}\mathbf{N}\mathbf{1} = \begin{bmatrix}\frac{1}{4}&\frac{1}{2}&\frac{1}{4}\end{bmatrix}\begin{bmatrix}\frac{13}{10}&\frac{2}{5}&\frac{1}{10}\\\frac{6}{5}&\frac{8}{5}&\frac{2}{5}\\\frac{9}{10}&\frac{6}{5}&\frac{13}{10}\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix} = \begin{bmatrix}\frac{1}{4}&\frac{1}{2}&\frac{1}{4}\end{bmatrix}\begin{bmatrix}\frac{9}{5}\\\frac{16}{5}\\\frac{17}{5}\end{bmatrix} = \frac{29}{10} = 2.9.$$

That is, the mean number of steps to absorption is 2.9. The second moment of the absorption time is

$$\mathrm{E}\left[W^2\right] = \mathbf{p}_T^{(0)}(2\mathbf{N}-\mathbf{I})\mathbf{N}\mathbf{1} = \begin{bmatrix}\frac{1}{4}&\frac{1}{2}&\frac{1}{4}\end{bmatrix}\begin{bmatrix}\frac{8}{5}&\frac{4}{5}&\frac{1}{5}\\\frac{12}{5}&\frac{11}{5}&\frac{4}{5}\\\frac{9}{5}&\frac{12}{5}&\frac{8}{5}\end{bmatrix}\begin{bmatrix}\frac{9}{5}\\\frac{16}{5}\\\frac{17}{5}\end{bmatrix} = \frac{1266}{100}.$$

Therefore, the variance is

$$\mathrm{Var}(W) = \mathrm{E}\left[W^2\right] - (\mathrm{E}[W])^2 = \frac{1266}{100} - \frac{841}{100} = \frac{425}{100} = \frac{17}{4}.$$ □

Note that the mean and variance of the time to absorption is influenced by the initial transition state vector. Intuitively, this should make sense – if the initial state distribution makes it more likely to start out in a state which is more likely to lead to an absorbing state, we should expect our expected time to absorption to decrease. Table 9.3.4 uses Equations (9.3.7), (9.3.9) and (9.3.10)

Let m_j be the expected (mean) absorption time (number of steps) from a transient state j to one of the absorbing states. Let $\mathbf{M}$ be a column vector whose components are m_j. From Equations (9.3.7) and (9.3.8), it follows that

$$\mathbf{M} = \mathbf{N}\mathbf{1}. \tag{9.3.11}$$

Let the second moment of the absorption time from the transient state j to the set of absorbing states be denoted by $m_j^{(2)}$. From equations (9.3.9) and (9.3.11), it follows that

$$\mathbf{M}^{(2)} = \{m_j^{(2)}\} = (2\mathbf{N}-\mathbf{I})\mathbf{M}.$$

Initial transition state vector $\mathbf{p}_T^{(0)}$	Mean absorption time $\mathrm{E}[W]$	Variance of absorption time $\mathrm{Var}(W)$
$\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$	1.80	2.8800
$\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$	3.20	3.8400
$\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$	3.40	4.8000
$\begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$	2.55	4.1675
$\begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}$	2.95	4.5275
$\begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$	2.90	4.2500

Table 9.1: Properties of the distribution of absorption times for the random walk of Example 9.3.9 using various initial probability vectors $\mathbf{p}_T^{(0)}$.

Let the variance of the absorption time for a transient state j be denoted by v_j. Since the variance equals the second moment minus the square of the mean,

$$\mathbf{V} = \{v_j\} = \mathbf{M}^{(2)} - \mathbf{M}_{sq} = (2\mathbf{N} - \mathbf{I})\mathbf{M} - \mathbf{M}_{sq}$$

where $\mathbf{M}_{sq} \equiv \{m_j^2\}$. That is, each element of the column matrix $\mathbf{M}$ is squared.

Example 9.3.12. As a continuation of Example 9.3.11, recall that

$$\mathbf{N} = \begin{bmatrix} \frac{13}{10} & \frac{2}{5} & \frac{1}{10} \\ \frac{6}{5} & \frac{8}{5} & \frac{2}{5} \\ \frac{9}{10} & \frac{6}{5} & \frac{13}{10} \end{bmatrix}.$$

Therefore, the column of expected number of steps to absorption is

$$\mathbf{M} = \mathbf{N}\mathbf{1} = \begin{bmatrix} \frac{13}{10} & \frac{2}{5} & \frac{1}{10} \\ \frac{6}{5} & \frac{8}{5} & \frac{2}{5} \\ \frac{9}{10} & \frac{6}{5} & \frac{13}{10} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{9}{5} \\ \frac{16}{5} \\ \frac{17}{5} \end{bmatrix} = \begin{bmatrix} 1.8 \\ 3.2 \\ 3.4 \end{bmatrix}.$$

For instance, the expected number of steps from state 3 to absorption is 3.4. To find the second moments,

$$\mathbf{M}^{(2)} = (2\mathbf{N} - \mathbf{I})\mathbf{M} = \begin{bmatrix} \frac{8}{5} & \frac{4}{5} & \frac{1}{5} \\ \frac{12}{5} & \frac{11}{5} & \frac{4}{5} \\ \frac{9}{5} & \frac{12}{5} & \frac{8}{5} \end{bmatrix} \begin{bmatrix} \frac{9}{5} \\ \frac{16}{5} \\ \frac{17}{5} \end{bmatrix} = \begin{bmatrix} \frac{153}{25} \\ \frac{352}{25} \\ \frac{409}{25} \end{bmatrix}.$$

Therefore, the column of variances of expected steps to absorption is

$$\mathbf{V} = \mathbf{M}^{(2)} - \mathbf{M}_{sq} = \begin{bmatrix} \frac{153}{25} \\ \frac{352}{25} \\ \frac{409}{25} \end{bmatrix} - \begin{bmatrix} \frac{81}{25} \\ \frac{256}{25} \\ \frac{289}{25} \end{bmatrix} = \begin{bmatrix} \frac{72}{25} \\ \frac{96}{25} \\ \frac{120}{25} \end{bmatrix}.$$

From this vector, we see that the variance in absorption time is the greatest for state 4 and the least for state 2. □

9.3.5 Limiting Distributions for Finite Markov Chains

the finite markov chains discussed in this section will have r states. If k is a transient state, then for any state j, $\lim_{n\to\infty} p_{jk}^{(n)} = 0$. This follows from the fact that for a transient state k,

$$\sum_{n=1}^{\infty} p_{jk}^{(n)} < \infty \quad \forall j \in S.$$

If k is not a transient state, then the problem of finding $\lim_{n\to\infty} p_{jk}^{(n)}$, if it exists, is in general more difficult.

Definition 9.3.8. A finite Markov chain is *ergodic* if there exists probabilities π_j such that

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_k \quad \forall i, j \in S. \tag{9.3.12}$$

These *limiting probabilities* $\{\pi_j\}$ are the probabilities of being in a state after equilibrium has been achieved.

As can be seen from (9.3.12), the values $\{\pi_j\}$ are independent of the initial state i. in fact, we have the following theorem about these unconditional probabilities:

Theorem 9.3.1. If $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$, then $\lim_{n\to\infty} p_j^{(n)} = \pi_j$.

Proof: By definition (or the Chapman-Kolmogorov equations),

$$p_j^{(n)} = \sum_{k=1}^{r} p_k^{(0)} p_{kj}^{(n)}.$$

Therefore,

$$\lim_{n\to\infty} p_j^{(n)} = \lim_{n\to\infty} \sum_{k=1}^{r} p_k^{(0)} p_{kj}^{(n)} = \sum_{k=1}^{r} p_k^{(0)} \lim_{n\to\infty} p_{kj}^{(n)} = \sum_{k=1}^{r} p_k^{(0)} \pi_j = \pi_j \sum_{k=1}^{r} p_k^{(0)} = \pi_j \cdot 1 = \pi_j.$$

■

The limiting probabilities $\{\pi_j\}$ may be found by solving the following system of equations, derived from the Chapman-Kolmogorov equations:

$$\pi_j = \sum_{k=1}^{r} \pi_k p_{kj}, \quad j \in \{1, \ldots, r\} \tag{9.3.13}$$

subject to the conditions

$$\pi_j \geq 0 \ \ \forall j; \qquad \sum_{j=1}^{r} \pi_j = 1. \tag{9.3.14}$$

The probability distribution $\{\pi_j\}$ defined by (9.3.13) and (9.3.14) is called a *stationary distribution.*

If a Markov chain is ergodic, then it can be shown (not here) that it possesses a unique stationary distribution. That is, there exists one and only one set of numbers $\{\pi_j\}$ that satisfy Equations (9.3.12), (9.3.13) and (9.3.14). However, there are Markov chains that have stationary distributions (which satisfy equations (9.3.12), (9.3.13) and (9.3.14)) which are not ergodic. For example, consider the tpm

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Then

$$p_{11}^{(n)} = \begin{cases} 1 & \text{if } n \text{ is even} \\ 0 & \text{if } n \text{ is odd} \end{cases},$$

so (9.3.12) is not satisfied, and so the chain is not ergodic. However, we can solve (9.3.13) and (9.3.14) to get stationary probabilities $\pi_1 = \pi_2 = \frac{1}{2}$. recall that this $\mathbf{P}$ is the transition matrix for an irreducible periodic Markov chain.

the following theorems will be stated without proof. They give some sufficient conditions for a finite Markov chain to be ergodic.

Theorem 9.3.2. A finite irreducible aperiodic Markov chain is ergodic.

Example 9.3.13. Let

$$\mathbf{P} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ 0 & \frac{2}{3} & \frac{1}{3} \\ \frac{3}{4} & \frac{1}{4} & 0 \end{bmatrix}.$$

This chain is irreducible (all pairs communicate), since it can be shown that $p_{ij}^{(2)} > 0 \ \forall i, j \in S$. It is aperiodic, since $p_{11} = \frac{1}{4} > 0$. Hence by theorem 9.3.2, this chain is ergodic. To find the limiting distributions, solve (9.3.13) to obtain

$$\begin{aligned} \pi_1 &= \frac{1}{4} \cdot \pi_1 + \frac{2}{3} \cdot \pi_3 \\ \pi_2 &= \frac{1}{4} \cdot \pi_1 + \frac{2}{3} \cdot \pi_2 + \frac{1}{4} \cdot \pi_3 \qquad \Rightarrow \qquad \boldsymbol{\pi} = \begin{bmatrix} \frac{2}{7} & \frac{3}{7} & \frac{2}{7} \end{bmatrix}. \\ \pi_3 &= \frac{1}{2} \cdot \pi_1 + \frac{1}{3} \cdot \pi_2 \end{aligned}$$

Therefore, the asymptotic probability of being in state 0 is $\frac{2}{7}$, in state 1 is $\frac{3}{7}$, and in state 2 is $\frac{2}{7}$. □

Example 9.3.14. For the Ehrenfest diffusion model in Example 9.2.5, it can be shown that if the number of molecules k distributed between the two containers A and B is large, then the stationary probability $\pi_{k/2}$ for state $E_{k/2}$ is approximately one. The value of the stationary probability $\pi_{k/2}$ depends on k, and it can be shown (e.g., Feller (1968, p. 397)) that $\lim_{k \to \infty} \pi_{k/2} = 1$. For example, if $k = 1{,}000{,}000$, then the probability of finding more than 505,000 molecules in A is about 10^{-23}. □

Definition 9.3.9. A tpm is *doubly stochastic* if each column sums to 1. That is, if

$$\sum_{i=1}^{r} p_{ij} = 1 \quad \forall j \in S.$$

Recall that a tpm must by definition have each row sum to 1.

Theorem 9.3.3. If the tpm $\mathbf{P}$ for a finite irreducible aperiodic Markov chain with r states is doubly stochastic, then the stationary probabilities are given by

$$\pi_k = \frac{1}{r} \quad \forall k \in \{0, \ldots, r-1\}.$$

Example 9.3.15. The tpm

$$\mathbf{P} = \begin{bmatrix} \frac{1}{8} & \frac{1}{4} & \frac{1}{8} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \\ \frac{3}{8} & \frac{1}{4} & \frac{1}{4} & \frac{1}{8} \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$$

is irreducible, aperiodic and doubly stochastic. Therefore

$$\pi_k = \frac{1}{4} \quad \forall k \in \{0, 1, 2, 3\}.$$ □

9.3.6 Summary

A Markov chain is a probabilistic model for a physical system from which we have a sample $(X_0, X_1, \ldots, X_n)$. It is characterized by the equation

$$\mathrm{P}(X_k | X_1, \ldots, X_{k-1}) = \mathrm{P}(X_k | X_{k-1}) \quad \forall k \in \mathbb{Z}^+.$$

That is, the state of the system at any give time depends only on the previous state.

A Markov chain is completely described when the state space S, the initial probability vector $\mathbf{p}^{(0)}$, and the one-step transition probabiliy matrix $\mathbf{P}$ are given. Using the information contained in S, $\mathbf{p}^{(0)}$ and $\mathbf{P}$, it is possible to compute the following:

1. The probability of going from state j to state k in n steps (time intervals). That is,

   $$p_{jk}^{(n)} = \sum_{i \in S} p_{ji} p_{ik}^{(n-1)} \quad \forall j, k \in S.$$

   This function may be written in matrix form as

   $$\mathbf{P}^{(n)} = \mathbf{P}^n = \mathbf{P}\mathbf{P}^{n-1},$$

   where $\mathbf{P}^{(n)}$ is called the n-step transition probability matrix.

2. The unconditional probability of being in state j at time n (n steps after the first observation). that is,

   $$p_j^{(n)} = \sum_{i \in S} p_i^{(0)} p_{ij}^{(n)} \quad \forall j \in S.$$

   This function may be written in matrix form as

   $$\mathbf{p}^{(n)} = \mathbf{p}^{(0)} \mathbf{P}^n.$$

3. The probability that the time W to absorption in a finite absorbing Markov chain. That is,

   $$\mathrm{P}(W = n) = \mathbf{p}_T^{(0)} \mathbf{Q}^{n-1} \mathbf{R}\mathbf{1}.$$

   The mean (expected) time to absorption, the second moment, and the variance may also be computed.

4. For irreducible aperiodic finite markov chains, the limiting probability of being in a state j. That is,

   $$\lim_{n \to \infty} \mathrm{P}(X_n = E_j) = \lim_{n \to \infty} p_j^{(n)} = \pi_j.$$

10 Supplementary Problems

10.1 Algebra of Sets

1. To join a certain club, a person must be a lawyer, a liar, or both. There are 25 members in this club, of which 19 are lawyers and 16 are liars. How many club members are both lawyers and liars?

10.2 Fundamental Definitions and Axioms

1. Engine blocks coming off an assembly line are numbered serially. During one particular work shift, the six blocks produced are numbered 17850 through 17855. An inspector selects two of the blocks at random to test for stress damage. Write out the sample space of all possible pairs of blocks that might be inspected.
2. An urn contains six chips numbered 1 through 6. Three are drawn out without replacement. What outcomes are in the event "second smallest chip is a 3"?
3. To determine an *odd person out*, $m = 2k + 1$ players each toss a coin. If one player's coin turns up differently than all the others, that person is declared the odd person out. Suppose that $m = 3$ players play this game.
   (a) List the outcomes in the sample space S.
   (b) Let E be the event that no one is declared an odd person out. Which outcomes are in E?
   (c) Can you see a pattern in your answer to part (a) that would suggest, without enumeration, the number of outcomes in S if $m = 7$? What is that number?
4. Let A and B be any two events defined on a sample space S. Then S is the union of A, $A^c \cap B$, and what other mutually exclusive event?
5. Winthrop, a premed student, has been summarily rejected by all 126 US medical schools. Desperate, he sends his transcripts and MCATs to the two least selective foreign schools he can think of, the two branch campuses (X and Y) of Swampwater Tech. Based on the successes his friends have had there, he estimates that his probability of being accepted at X (event A) is 0.7 and at Y (event B) is 0.4. He also suspects he knows the chance that at least one of the applications will be rejected. Compute $\mathrm{P}(A \cup B)$ if
   (a) $\mathrm{P}(A^c \cup B^c) = 0.7$.
   (b) $\mathrm{P}(A^c \cup B^c) = 0.9$.
6. An experiment has two possible outcomes: The first occurs with probability p and the second occurs with probability p^2. Find p.
7. Suppose S has a finite number of outcomes, all equally likely. Let $n(A)$ denote the number of outcomes in the event A, where $A \subseteq S$. Define
$$\mathrm{P}(A) = \frac{n(A)}{n(S)}.$$
   Show that $\mathrm{P}(A)$ satisfies Axioms 1 through 3, as explained in Definition 2.3.1.
8. (a) Specify a sample space for the experiment that consists of drawing 1 ball from an urn containing 10 balls, of which 4 are white and 6 are red, assuming that the balls are numbered 1 through 10.
   (b) Specify a sample space for the experiment that consists of drawing 2 balls with replacement from the urn containing 10 balls (that is, the first ball removed is placed in the urn before the second is drawn out). Assume that order matters - getting a 1 and then a 2 is different than getting a 2 and then a 1. Again assume that they are colored and numbered as in 8(a).

(c) Specify a sample space for the experiment that consists of drawing 2 balls without replacement from the urn containing 10 balls (that is, the first ball removed is not placed in the urn before the second is drawn out). Assume that order matters - getting a 1 and then a 2 is different than getting a 2 and then a 1. Again assume that they are colored and numbered as in 8(a).

(d) For the sample space given in 8(a), define the events (as subsets):

i. A: A white ball is drawn.

ii. B: A red ball is drawn.

(e) For the sample space given in 8(b), define the events (as subsets):

i. C: The first ball is white.

ii. D: The second ball is white.

iii. E: Both balls are white.

Does $C \cap D = E$?

9. Two light bulbs are placed on a test until both fail. Assume that each will burn no more than 1600 hours. Define a reasonable sample space for this experiment and describe, as subsets, the events:

(a) A: Both bulbs fail in less than 1000 hours.

(b) B: Neither bulb fails in less than 1000 hours.

(c) C: The shorter time to failure of the two is 1000 hours.

(d) D: The longer time to failure of the two is 1000 hours.

10. Given $S = \{1, 2, 3\}$, $A = \{1\}$, $B = \{3\}$, $C = \{2\}$, $\mathrm{P}(A) = \frac{1}{3}$ and $\mathrm{P}(B) = \frac{1}{3}$, find

(a) $\mathrm{P}(C)$.

(b) $\mathrm{P}(A \cup B)$.

(c) $\mathrm{P}(A^c)$.

(d) $\mathrm{P}(A^c \cap B^c)$.

(e) $\mathrm{P}(A^c \cup B^c)$.

(f) $\mathrm{P}(B \cup C)$.

11. Prove, from the three axioms, that $\mathrm{P}(A) \leq 1$ for every set A.

12. Is it possible to have an assignment of probabilities such that $\mathrm{P}(A) = \frac{1}{2}$, $\mathrm{P}(A \cap B) = \frac{1}{3}$ and $\mathrm{P}(B) = \frac{1}{4}$? Why?

13. How many 5-man squads could be chosen from a company of 20 men?

14. A university committee has 100 members, 60 of whom favor giving Nate Derby a huge raise in his salary. The committee president is going to randomly choose 10 members (i.e., each has the same chance of being chosen) who will then vote on whether to give Nate that raise.

(a) How many different sets of 10 people can be made from the 100 members?

(b) How many of these sets will include 6 or more people who favor giving Nate that raise?

(c) How many of these sets will not include 6 or more people who favor giving Nate that raise?

(d) If a majority is needed for it (i.e., a tie won't do it), what is the probability that Nate will get that huge raise?

15. How many different 11-letter sequences can be made using the letters in the word MISSISSIPPI? How many of these begin with an M and end with an I?

16. Each state has 2 senators. What is the probability that in a committee of 50 senators chosen at random (so that each senator has the same probability of being selected),

(a) Washington State is represented?

(b) Each state is represented?

17. *Johnny Carson screws up the birthday problem.* Ross (2006, Ex. 5i, pp. 41-42) explains the birthday problem, where we see that if we have 23 or more people in the room, the probability that no two of them have the same birthday is less than 50%. According to one version of the story, Johnny Carson once had a guest on his show who mentioned this to him. Carson didn't believe it, so he asked his audience of about 120 people if anyone shared his birthday, October 23. No one did, and Carson remained unconvinced.

    (a) Why is Carson's question very different from the original birthday problem?

    (b) What is the actual probability that none of the 120 people in his audience shared his birthday?

    (c) What is the minimum number of people needed in the audience for the probability of none of them sharing his birthday to be below 50%?

18. If X is a discrete random variable, express $\mathrm{P}(X \geq a)$ in terms of the distribution function of X.

19. Suppose that the distribution function of X is given by

$$F_X(b) = \begin{cases} 0 & b < 0 \\ \frac{b}{4} & 0 \leq b < 1 \\ \frac{1}{2} + \frac{b-1}{4} & 1 \leq b < 2 \\ \frac{11}{12} & 2 \leq b < 3 \\ 1 & 3 \leq b \end{cases}.$$

    (a) Find $\mathrm{P}(X = i)$ for $i \in \{1, 2, 3\}$.

    (b) Find $\mathrm{P}(\frac{1}{2} < X < \frac{3}{2})$.

20. If the distribution function of X is given by

$$F_X(b) = \begin{cases} 0 & b < -1 \\ \frac{1}{4} & -1 \leq b < 1 \\ \frac{3}{5} & 1 \leq b < 2 \\ \frac{4}{5} & 2 \leq b < 4 \\ \frac{9}{10} & 4 \leq b < 4.5 \\ 1 & b \geq 4.5 \end{cases},$$

    then calculate the probability mass function of X.

21. Nate can remain sane while teaching an accelerated class for a random amount of time X, after which he loses his mind. If the density of X is given (in units of weeks) by

$$f_X(x) = \begin{cases} Cxe^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases},$$

    what is the probability that Nate remains sane for the 6 weeks left in the summer term?

22. The pdf of X, the lifetime of a new Apple iPhone 3G (measured in months), is given by

$$f_X(x) = \begin{cases} \frac{10}{x^2} & x > 10 \\ 0 & x \leq 10 \end{cases}.$$

(a) Find $P(X > 20)$.

(b) What is the cumulative distribution function of X?

(c) Assuming that the reliability of each iPhone is independent of any other one, what is the probability that of 6 iPhones, at least 3 of them will function for at least 15 months?

23. Let

$$f_{XY}(x,y) = \begin{cases} 21x^2y^3 & 0 < x < y < 1 \\ 0 & \text{otherwise} \end{cases}.$$

Find

(a) $f_X(x)$.

(b) $f_Y(y)$.

(c) $F_{XY}(\frac{1}{2}, \frac{2}{3})$.

(d) $F_X(\frac{1}{2})$.

Are X and Y independent? Why?

24. (a) A student takes a true-false exam that has four questions: Assume she is guessing at the right answer to each question. Define X_1 = the number she gets right of the first two questions, X_2 = the number she gets right of the last two questions.

i. Derive the pmf for (X_1, X_2).

ii. Repeat this exercise assuming each exam question is a multiple choice with four possible responses.

(b) For the case discussed above, define Y_1 = the number she gets right of the first 3 questions and Y_2 = the number she gets right of the last 3 questions. Answer the above two questions for (Y_1, Y_2).

(c) What are the marginal pmfs for X_1, X_2, Y_1, Y_2 defined above?

25. What must A equal if

$$f_{XY}(x,y) = \begin{cases} A\dfrac{x}{y} & x \in (0,1),\ y \in (1,2) \\ 0 & \text{otherwise} \end{cases}$$

is to be a density function? Are X and Y independent? Why?

26. Suppose that X and Y are continuous random variables with the joint pdf

$$f_{XY}(x,y) = \begin{cases} e^{-y} & x \in (0,1),\ y > 0 \\ 0 & \text{otherwise} \end{cases}.$$

Find the marginal pdfs is X and Y. Are X and Y independent?

27. Assume (X, Y) has the density

$$f_{XY}(x,y) = \frac{1}{2}$$

for (x, y) inside the square with corners (a, a), $(a, -a)$, $(-a, a)$, and $(-a, -a)$, and that $f_{XY}(x, y)$ is zero elsewhere.

(a) Find a.

(b) Find the marginal densities for X and Y.

(c) Are X and Y independent? Why?

10.3 Dependent and Independent Events

1. If $\mathrm{P}(A) = a$ and $\mathrm{P}(B) = b$, then show that $\mathrm{P}(A|B) \geq \frac{a+b-1}{b}$.

2. Consider a sample size of 3 balls drawn in the following manner: We start with an urn containing 5 white and 7 red balls. At each trial a ball is drawn and its color is noted. The ball drawn is then returned to the urn, together with an additional ball of the same color. Find the probability that the sample will contain exactly

   (a) 0 white balls.

   (b) 1 white ball.

   (c) 3 white balls.

3. In answering a question on a multiple-choice test, a student either knows the answer or guesses. Let p be the probability that the student knows the answer and $1-p$ the probability that the student guesses. Assume that a student who guesses at the answer will be correct with probability $\frac{1}{m}$, where m is the number of multiple-choice alternatives. What is the conditional probability that a student knew the answer to a question, given that he or she answered it correctly?

4. If two fair dice are rolled, what is the conditional probability that the first one lands on 6, given that the sum of the dice is i? Compute this for all values of i between 2 and 12.

5. Suppose that you continually collect coupons and that there are m different types of them. Suppose also that each time a new coupon is obtained, it is a type i coupon with probability p_i, $i \in \{1, \ldots, m\}$. Now suppose that you have just collected your nth coupon. What is the probability that it is a new type of coupon?

6. From a set of n randomly chosen people, let E_{ij} denote the event that persons i and j have the same birthday. Assume that each person is equally likely to have any one of the 365 days of the year as his/her birthday. Find

   (a) $\mathrm{P}(E_{3,4}|E_{1,2})$.

   (b) $\mathrm{P}(E_{1,3}|E_{1,2})$.

   (c) $\mathrm{P}(E_{2,3}|E_{1,2} \cap E_{1,3})$.

   What can you conclude from the above about the independence of the $\binom{n}{2}$ events E_{ij}?

7. A family has n children with probability $\alpha p^n, n \geq 1$, where $\alpha \leq \frac{1-p}{p}$.

   (a) What proportion of families have no children?

   (b) If each child is equally likely to be a boy or a girl (independently of each other), what proportion of families consists of i boys (and any number of girls)?

8. Consider an American roulette wheel with 38 slots: 18 red, 18 black, and 2 green. If you bet \$1 on red, you win \$1, and thus receive \$2 total. Otherwise, you lose your initial dollar. Kat has \$3 and is following a policy of betting \$1 on red each time. She will stop on either her first win or when she loses all her money. Let $X =$ Kat's gain. Find the pmf of X, as well as $\mathrm{E}[X]$.

10.4 Probability Laws

1. Independent trials that result in a success with probability p are successively performed until a total of r successes is obtained. Show that the probability that exactly n trials are required is

$$\binom{n-1}{r-1} p^r (1-p)^{n-r}.$$

2. Five fair dice are rolled. Let X denote the number of 1's that occur. Find

(a) $\mathrm{P}(1 \leq X \leq 4)$.

(b) $\mathrm{P}(X \geq 2)$.

3. Ten fair coins are tossed onto a table. Let Z denote the number of coins which land heads up. Compute $\mathrm{P}(Z = 5)$.

4. A bag contains 10 flashbulbs, 8 of which are good. If 5 flashbulbs are chosen from the bag at random (without replacement), what is the probability mass function for the number of good flashbulbs chosen? What is the probability mass function for the number of bad flashbulbs chosen?

5. $X \sim \mathrm{B}(n, p)$. What value of p maximizes $\mathrm{P}(X = k)$ for $k \in \{1, \ldots, n\}$? This is known as the *maximum likelihood estimation* method of estimating p. That is, if we assume that $X \sim \mathrm{B}(n, p)$ and n is known, then we estimate p by choosing the value of p that maximizes $\mathrm{P}(X = k)$.

6. Assume that 1 baby in 10,000 is born blind. If a large city hospital has 5000 births in a given year, approximate the probability that none of the babies born that year was blind at birth. Also approximate the probabilities that exactly 1 is born blind, and that at least 2 are born blind.

7. Let X be a Poisson random variable with parameter λ. What value of λ maximizes $\mathrm{P}(X = k),\ k \geq 0$? As with problem above, this is the *maximum likelihood estimate* of λ when we have data and want to make a best guess as to what λ is.

8. In the morning, students enter the STAT/MATH 394 class at a rate of 1 for every 3 minutes.

   (a) What is the probability that no one enters between 8:15 and 8:20?

   (b) What is the probability that at least 4 students enter the classroom during that time?

9. Calls arrive at a switchboard according to a Poisson process with parameter $\lambda = 5$ per hour. If we are at the switchboard, what is the probability that

   (a) it is at least 15 minutes until the next call?

   (b) it is no more than 10 minutes until the next call?

   (c) it is exactly 5 minutes until the next call?

10. A newsboy is selling papers on The Ave. The paper he sells are events in a Poisson process with parameter $\lambda = 50$ per hour. If we have just purchased a paper from him, what is the probability that it will be at least 2 minutes until he sells another? If it is already 5 minutes since his last sale, what is the probability that it will be at least 2 more minutes until his next sale?

11. Suppose that n independent trials are performed such that each one has a probability p of being a success. These trials are performed over and over again until a success occurs. If X equals the number of trials required, show that

$$\mathrm{P}(X = n) = (1 - p)^{n-1} p, \quad n \in \mathbb{Z}^+$$

    and that these probabilities add up to one over all values of n. This is called the *geometric distribution.*

12. It has been assumed empirically that deaths per hour due to traffic accidents occur at a rate of 8 per hour on July 4th weekend in the US. Assuming that these deaths occur independently under a Poisson process, compute the probabilities that

    (a) A 1-hour period would pass with no deaths.

    (b) A 15-minute interval would occur with no deaths.

    (c) 4 consecutive 15-minute periods would occur with no deaths.

13. X is uniformly distributed on $(0, 2)$ and Y is exponential with parameter λ. Find the value of λ such that $\mathrm{P}(X < 1) = \mathrm{P}(Y < 1)$.

14. One student loves STAT/MATH 394 A so much, the time that he arrives to class is a normal random variable with $\mu = 8{:}05$ and $\sigma = 8$ minutes. What is the probability that

    (a) He arrives before 8:00?

    (b) He arrives between 8:15 and 8:30?

    (c) He arrives late for class?

    When doing this problem, for convenience feel free to convert the problem from $X = \{\text{time that student arrives}\}$ to $Y = \{\text{minutes after 8:00 that student arrives}\}$.

10.5 Functions of Random Variables

1. Let X have the pmf

$$p_X(x) = \begin{cases} \frac{1}{8} & x \in \{\pm 1, \pm \frac{1}{2}\} \\ \frac{1}{2} & x = 0 \\ 0 & \text{otherwise} \end{cases}.$$

    Find the pmf for the following functions:

    (a) X^2.

    (b) e^X.

    (c) $2X + 1$.

    (d) $2X^2 + 1$.

2. Suppose that a fair six-sided die is rolled twice. What are the possible values and associated probabilities that the following random variables can take on?:

    (a) The maximum value to appear in the two rolls.

    (b) The minimum value to appear in the two rolls.

    (c) The sum of the two rolls.

    (d) The value of the first roll minus the value of the second roll.

3. Let X be uniformly distributed on the interval from -1 to 1. Find the pdf of the functions given:

    (a) X^2.

    (b) e^X.

    (c) $2X + 1$.

    (d) $2X^2 + 1$.

4. Let X be normally distributed with parameters $\mu = 0$ and $\sigma^2 = 1$. Find the pdf of the functions given:

    (a) X^2.

    (b) e^X.

    (c) $2X + 1$.

5. Find the pdf of $X = \cos(\theta)$, where θ is uniformly distributed from $-\pi$ to π.

6. Let $X_1, X_2 \overset{iid}{\sim} U(0, 1)$. Find the pdf of the functions given:

    (a) $X_1 + X_2$.

(b) $X_1 - X_2$.

(c) $\min(X_1, X_2)$.

(d) $\max(X_1, X_2)$.

(e) $X_1 X_2$.

7. Let $X_1, X_2 \overset{iid}{\sim} \text{Exp}(\lambda)$. Find the pdf of $Y = X_1 - X_2$ using Jacobians.

8. Let $X_1, X_2 \overset{iid}{\sim} \text{P}(\lambda)$. Find the pmf of $Y = X_1 + X_2$.

9. For a given $n \in \mathbb{Z}^+ = \{1, 2, 3, \ldots\}$, a *Chi-squared distribution with n degrees of freedom*, denoted χ^2_n, is defined as a $\text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$ distribution. Suppose $X \sim \chi^2_n$.

   (a) What is the pdf of X?

   (b) What is the pdf of $Y = \frac{X}{1+X}$?

   (c) It can be shown that if $Y_1 \sim \chi^2_{n_1}$, $Y_2 \sim \chi^2_{n_2}$, and the two of them are independent, then $Y_1 + Y_2 \sim \chi^2_{n_1+n_2}$. Use this to prove by induction that if $Z_1, Z_2, \ldots, Z_n \overset{iid}{\sim} \text{N}(0,1)$, then $\sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$.

10.6 Mathematical Expectation

1. Let X have the pmf

$$p_X(x) = \begin{cases} \frac{1}{5} & x \in \{2, 4, 6, 8, 16\} \\ 0 & \text{otherwise} \end{cases}.$$

   Calculate the following:

   (a) $\text{E}[X]$.

   (b) $\text{E}[X^2]$.

   (c) $\text{E}\left[\frac{1}{X}\right]$.

   (d) $\text{E}\left[2^{X/2}\right]$.

   (e) σ^2_X.

   (f) σ_X.

2. If $\text{E}[X] = 1$ and $\text{Var}(X) = 5$, then find

   (a) $\text{E}[(2+X)^2]$.

   (b) $\text{Var}(4 + 3X)$.

3. Let X be a binomial random variable with parameters n and p, denoted as $X \sim \text{B}(n, p)$. Show that

$$\text{E}\left[\frac{1}{X+1}\right] = \frac{1-(1-p)^{n+1}}{(n+1)p}.$$

4. Suppose there is a lottery with 10,000 tickets and a grand prize of \$3000. What is the expected value and variance of your winnings if a ticket costs \$1 and

   (a) you buy one lottery ticket?

   (b) you buy 100 lottery tickets?

5. *How insurance companies make their money.* An insurance company writes a policy to the effect that an amount of money m must be paid if the policy holder dies within one year. If a smart student from STAT/MATH 394 who then went on to pass the first exam in the fall of 2008 estimates that the policy holder will die within one year with probability p, what should the insurance company charge for its expected profit to be 15% of m?

6. *The St. Petersburg Paradox* A person tosses a fair coin until a tail appears for the first time. If the tail first appears on the nth flip, the person wins 2^n dollars. Let N denote the flip that results in the first tail.

   (a) Find $\mathrm{P}(N = n)$ for $n \in \mathbb{Z}^+$.

   (b) What is the expected value of the player's winnings?

   (c) A game is considered *fair* if the expected value of the earnings equals \$0. How much money should someone pay to play this game if the game is to be fair?

7. Now suppose the game described in problem 6 is truncated at 20 flips. That is, the game stops on the 20th flip if it hasn't stopped already.

   (a) What is the expected value of the player's winnings?

   (b) How much money should someone pay to play this game if the game is to be fair?

8. Now suppose the game described in problem 6 is modified such that if $N > 20$ (i.e., if the game continues beyond the 20^{th} flip), you win $\$(2^{20})$.

   (a) What is the expected value of the player's winnings?

   (b) How much money should someone pay to play this game if the game is to be fair?

9. Suppose the pdf of X is given by

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

   Calculate the following:

   (a) $\mathrm{E}[X]$.

   (b) $\mathrm{E}[X^2]$.

   (c) $\mathrm{E}[(X+10)^2]$.

   (d) $\mathrm{E}\left[\frac{1}{1-X}\right]$.

   (e) $\mathrm{Var}(X)$.

   (f) $\mathrm{SD}(X)$.

10. Show that

$$\mathrm{E}[Y] = \int_0^\infty \mathrm{P}(Y > y)dy - \int_0^\infty \mathrm{P}(Y < -y)dy.$$

11. (a) A fire station is to be located along a road of length A, $A < \infty$. If fires will occur at points uniformly distributed over the interval $(0, A)$, where should the station be located so as to minimize the expected distance from the fire? That is, choose a so as to minimize $\mathrm{E}[|X - a|]$ when X is uniformly distributed over $(0, A)$.

    (b) Now suppose that the road is of infinite length, stretching from point 0 outwards toward ∞. If the distance of a fire from point 0 is exponentially distributed with rate λ, where should the fire station now be located? That is, choose a so as to minimize $\mathrm{E}[|X - a|]$ when X is now exponentially distributed with rate λ.

12. Let $X \sim \mathrm{N}(1, \sigma^2 = 4)$ be independent of $Y \sim \mathrm{N}(1, \sigma^2 = 9)$. Find the following:

(a) $\mathrm{E}[3X - 9Y + 4]$.

(b) $\mathrm{E}\left[X^2 + Y^2\right]$.

(c) $\mathrm{Var}(3X - 9Y + 4)$.

(d) $\mathrm{P}(X < Y)$. [Hint: What is the distribution of $X - Y$?]

(e) the pdf of $V = 3X - 9Y + 4$.

(f) the pdf of $W = \frac{(3X-9Y+6)^2}{765}$.

13. N people arrive separately to a professional dinner. Upon arrival, each person looks to see if he or she has any friends among those present. That person either sets at the table of a friend or at an unoccupied table if none of those present is a friend. Assuming that each of the $\binom{N}{2}$ pairs of people are, independently, friends with probability p, and that there are N tables available, find the expected number of occupied tables after all N people arrive.

    Hint: Let X_i equal 1 or 0, depending on whether the ith arrival sits at a previously unoccupied table.

14. Let $X \sim \mathrm{U}(a, b)$. Find its moment generating function and characteristic function.

15. Let $X \sim \text{Geometric}(p)$. Find its moment generating function and use it to show that $\mathrm{E}[X] = \frac{1}{p}$.

    Hint: For this solution, you probably need to use the formula for the sum of a geometric series – i.e., $\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$ for $|r| < 1$. But how do we know that $|r| < 1$ in this problem? (You might need to use some real analysis for this) Alternatively, can you find a way to correctly do this problem *without* using the formula for the sum of a geometric series?

16. Use moment generating functions to prove that if $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$, then $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

    The moment generating function for a $\mathrm{N}(\mu, \sigma^2)$ distribution is $e^{\mu t + \sigma^2 t^2/2}$. You can use this for this problem, or you can prove that this is the correct moment generating function.

17. In Homework 2, Problem 1(c), we assumed that if $Y_1 \sim \chi^2_{n_1}$, $Y_2 \sim \chi^2_{n_2}$, and the two of them are independent, then $Y_1 + Y_2 \sim \chi^2_{n_1+n_2}$. Now prove it, using moment generating functions.

    The moment generating function for a χ^2_n distribution is $\left(\frac{1}{1-2t}\right)^{n/2}$. You can use this for this problem, or **for extra credit**, you can prove that this is the correct moment generating function.

10.7 Jointly Distributed Random Variables, Continued

1. Use the logic from class handout 2 to show that if $X_1, \ldots X_n \overset{iid}{\sim} f_X(x)$, then

$$f_{X_{(i)}X_{(j)}}(x_i, x_j) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \left[F_X(x_i)\right]^{i-1} \left[F_X(x_j) - F_X(x_i)\right]^{j-i-1} \left[1 - F_X(x_j)\right]^{n-j} f_X(x_i) f_X(x_j)$$

    A rigorous mathematical proof is not necessary; an informal argument will suffice.

2. Let $X_1, X_2, \ldots, X_{10} \overset{iid}{\sim} \mathrm{Exp}(\lambda)$. Find

    (a) the pdf of $X_{(1)}$.

    (b) the pdf of $X_{(10)}$.

    (c) the joint pdf of $X_{(1)}$ and $X_{(10)}$.

    (d) the pdf for the range $R = X_{(10)} - X_{(1)}$.

(e) $\mathrm{P}(R > 100 \text{ hours})$ if $\lambda = 0.01$ per hour

3. Suppose that X and Y are two random variables jointly distributed with pdf

$$f_{XY}(x,y) = \begin{cases} y^2 e^{-y(x+1)} & x, y \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

Find the following:

(a) $F_Y(y)$.
(b) $F_{XY}(2,4)$.
(c) $f_{X|Y}(x|y)$.
(d) $\mathrm{E}[X|Y=10]$.

4. Let X_1, X_2, X_3 be a random sample from the distribution with a pdf of $f_X(x) = 2x$ for $x \in (0,1)$. Find the probability that the smallest of these X_is exceeds the median of the distribution.

5. Let $X_1, X_2 \overset{iid}{\sim} N(0, \sigma^2)$. Show that $\mathrm{E}\left[X_{(1)}\right] = -\sigma/\sqrt{\pi}$.

6. Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathrm{U}(0,1)$. Show that

$$\mathrm{P}(X_{(k)} - X_{(k-1)} > t) = (1-t)^n, \qquad k \in \{1, 2, \ldots, n+1\},$$

where $X_0 \equiv 0$ and $X_{n+1} \equiv 1$.

10.8 Limiting Distributions

1. Assume that bags of Victrola coffee beans are packaged into boxes containing 50 bags, and that the expected weight is 16 ounces and the standard deviation is $\frac{1}{2}$ ounce. Find the approximate probability that a box of coffee beans (in bags) weighs less than 805 ounces.

2. Suppose that the number of airplanes arriving in any 30-minute period obeys a Poisson probability law with mean 64. Use Chebyshev's inequality to determine a lower bound for the probability that the number of airplanes arriving in a 30-minute period will be between 48 and 80, *non*-inclusive.

3. Two types of coins are produced at a factory: A fair coin and a biased one that comes up heads 55% of the time. We have one of these coins but do not know whether it is a fair coin or a biased one. In order to ascertain which type of coin we have, we shall perform the following *hypothesis test*: We shall toss the coin 1000 times. If the coin lands on heads 525 or more times, then we shall conclude that it is a biased coin; Otherwise, we shall conclude that it is fair. If the coin is actually fair, what is the probability that we shall reach a false conclusion (called a *type I error*)? What would it be if the coin were biased (called a *type II error*)?

4. From past experience, Garth Warner knows that the test score of a student taking his final exam is a random variable with mean 75.

(a) Give an upper bound for the probability that a student's test score will be 85 or above.

Suppose, in addition, that he knows that the variance of a student's test score is equal to 25.

(b) What can be said about the probability that a student will score between 65 and 85?
(c) How many students would have to take the exam so as to ensure, with probability at least .9, that the class average would be within 5 of 75? Use the weak law of large numbers.
(d) Use the central limit theorem to solve part (c) above.

(e) Of the two answers you have for (c) and (d) above, which method gave you the smaller answer (i.e., the number of students)? Why do you think that method gave you the smaller answer?

5. Upon graduating, Kelly Jones and Jamie Kam start a restaurant. Their marquee (optimally placed near an I-5 exit) uses a huge light bulb whose lifetime is exponentially distributed with mean of 5 hours.

   (a) If they have 100 of these light bulbs (whose lifetime distributions are independent from one another) and they are used one at a time, with a failed bulb being immediately replaced by a new one, what is the approximate probability that there is a still a working bulb after 525 hours?

   (b) Suppose that it takes a random time, uniformly distributed over $(0, .5)$ hours to replace a failed bulb. What is the approximate probability that there is still a working bulb after 550 hours?

6. (A problem from an actuarial exam) A local auto insurance company insures 100 individuals during a specific period. Suppose each insured claims independently, and the number of claims of each insured during this period has the following probability distribution:

X: x (possible claims)	0	1	2	3
$\mathrm{P}(X = x)$	0.64	0.14	0.10	0.12

   (a) What is the approximate probability that there are at least 70 total claims during this period?

   (b) What is the approximate probability that there are less than 60 insured individuals who claim nothing during this period?

7. In example 2.2 of page 93 of the notes, we see in the graph that the Gamma(n, λ) distribution can be approximated by the normal distribution. Use the Central Limit Theorem to prove that as $n \to \infty$, the Gamma(n, λ) distribution approaches the normal distribution. What will be the mean and variance of this normal distribution?

A Probability vs Statistics

According to one academic (Skiena, 2001),

> *Probability deals with predicting the likelihood of future events, while statistics involves the analysis of the frequency of past events.*

This is a generally accurate description, with one important distinction: While statistics analyzes past events, it is almost always done with the purpose of predicting the likelihood of future events. For example, if we test an experimental drug on lab mice, we don't really care about what happened to those mice - what we really care about is what will happen to future mice (and, eventually, future humans) if we administer the drug to them under the same conditions. Statistics thus can be divided into two categories:

- *Descriptive statistics* analyzes events of the past to answer the question "What happened?" without much thought of using that data to predict future events. This often involves nothing more than showing a few pie charts and graphs.
- *Inferential statistics* analyzes events of the past to answer the question "What is going to happen?". This involves building mathematical models for the data, and then using those models to answer the question.

As one might expect, the latter is far more difficult (and interesting) to do. Indeed, when you take a class in statistics at the university level, it is almost always inferential statistics and not descriptive statistics (which can be learned with little or no math).

Inferential statistics makes heavy use of probability theory to answer the question "What is going to happen?". Here is a concrete example:

- In probability, if you have a coin with probability p of getting heads and $1 - p$ of getting tails, you can figure out how to find the probability that flipping a coin 10 times will give you 5 heads.
- In descriptive statistics, if you take one coin and flip it many times, you can figure out how to come up with a fairly good estimate of p for the above problem.
- In inferential statistics, you can then use your estimate of p (and its estimated accuracy) to figure out the probability that if you flip this same coin 10 more times, that you will get 5 heads. Even more, you can come with a measure of accuracy for this prediction.

If it helps, we can think of it this way:

- Probability = predicting the likelihood of future events.
- Descriptive statistics = analyzing the frequency of past events.
- Inferential statistics = using the above two together to predict future events, given past events.

This might explain why a student can take probability without statistics, but not the other way around (without feeling completely overwhelmed).

Econometrics is a term that is commonly used to refer to those areas of (inferential) statistics that are used in economics - most notably, regression and time series analysis. Furthermore, it refers to areas in regression and time series that are used more by economists than by the statistical world in general - mostly for philosophical reasons. For example, the term *stochastic process* refers to any statistical process indexed over time or space. While statistics can involve any kind of stochastic process, econometrics often focuses on just a few kinds of stochastic processes, such as an autoregressive integrated moving average model.

Note that econometrics is not the only area of statistics used in economics: *Risk analysis* is a new area (ten years old or so) that involves looking at high-frequency data, such as in the stock market, using computational techniques that were not possible before the age of high-speed computers. This is not considered part of the word "econometrics" for mostly historic reasons: That word has been used for about 100 years only in reference to time series and regression, and it would be difficult to change that definition now.

B Counting: An Introduction

This is a tutorial on counting methods, intended to help anyone who has difficulty with counting methods used in Section 2.[1] The aim with these notes is to give a basic grounding in counting, and discuss the translation of word problems into the exact counting question desired.

B.1 Why Is Counting Important for Probability?

From the discussion on top of page 15, we know that if each point in the sample space S is equally likely, then

$$\mathrm{P}(A) = \frac{\text{Size of } A}{\text{Size of } S} = \frac{\text{Number of events in } A}{\text{Number of events in } S}. \tag{B.1.1}$$

It's usually very easy to compute the number of events in the sample space S. *For this special case of equally-likely events* (and only for this special case), this section gives techniques for determining the number of events in the event A, which then allows us to compute the fraction above.

B.2 How to Count

In this section, a counting problem will concern a set of objects and a subset with particular properties. The most important points in a counting problem in this course are the following:

1. Does order matter?
2. Are objects distinguishable from each other?
3. Are objects chosen with or without replacement?

In general, objects are chosen without replacement. If they are replaced, it is either obvious from context (i.e. how many different strings of letters can you get that are 10 letters long; here we obviously allow the same letter to be used more than once), or it is mentioned in the problem. If there is something stranger (in one problem, objects were chosen with new objects added after each choice depending on the choice made), then it will definitely be in the problem. I leave determining what to do with those situations as an exercise for the reader, but the basics shown here should provide enough guidance.

Example B.2.1. *A bridge hand has 13 cards. How many possible bridge hands are there?*

Here, order does not matter, only the particular cards in the hand do; every card is distinguishable from every other card; and they are chosen without replacement. □

Example B.2.2. *How many different ways are there to rearrange the letters of the word 'MISSISSIPPI'?*

Here order matters, there are distinguishable groups of indistinguishable objects (i.e. the letter M is different from I, but there is no difference between Ps), and the letters are chosen without replacement. □

[1] This appendix is written by D.J. Schreffler, who took STAT/MATH 394/395 from Mr. Derby in the summer of 2008. Ideas and problems in this appendix were derived from Bóna (2006) and D'Angelo and West (2000), as well as from Mr. Derby's homeworks.

Once the questions about order, distinguishability, and replacement are answered, the next questions are the following, which will set us up to compute (B.1.1):

1. What size is my sample space?
2. What size is the set of events that I am interested in?

Example B.2.3. *What is the probability of a full house in 5-card stud?*

Here we see that order does not matter, the objects are distinguishable, and they are chosen without replacement. What is the size of the sample space? The number of possible hands of 5 cards chosen from 52. How many of those are full houses? Those with a pair and a three-of-a-kind. □

Definition B.2.1. A *combination* is a subset of k distinguishable objects out of a set of n where $0 \leq k \leq n$. The standard example is a hand of cards. There are

$$\frac{n!}{k!(n-k)!}$$

combinations, written $\binom{n}{k}$ and pronounced "n choose k," taking $0! = 1$. From this, it should be obvious that $\binom{n}{k} = \binom{n}{n-k}$. This is also called the *binomial coefficient*, the reason for which I present below.

Example B.2.1 Revisited. *A bridge hand has 13 cards. How many possible bridge hands are there?*

There are 52 cards, and we choose 13 of them to be in a bridge hand, so we want $\binom{52}{13} = \frac{52!}{13!39!}$. □

Definition B.2.2. A *permutation* is an ordering of a set. Given a string of n different characters, there are $n!$ different ways that they can be ordered.

Example B.2.2 Revisited. *How many different ways are there to rearrange the letters of the word 'MISSISSIPPI'?*

Imagine that we label all of the duplicated letters as follows: $\mathrm{MI_1S_1S_2I_2S_3S_4I_3P_1P_2I_4}$. Then there are 11! different orderings of these characters. But all of the Is are identical, so divide by 4!. The Ss and Ps are identical, so divide by 4! and 2!. For completeness' sake, divide by 1! for the M to obtain

$$\frac{11!}{1!2!4!4!}.$$

This is also written as $\binom{11}{1,2,4,4}$. □

Definition B.2.3. Let $n = \sum_{i=1}^{k} a_i$, such that $a_i \geq 0 \ \forall i$. Then $\binom{n}{a_1,a_2,\ldots,a_k} = \frac{n!}{a_1!a_2!\ldots a_k!}$. These numbers are called *multinomial coefficients* for the same reason that $\binom{n}{k}$ is the binomial coefficient, and $\binom{n}{k}$ can be written as $\binom{n}{k,n-k}$ as well.

Regarding replacement: If objects are replaced, then it generally results in a power. (How many different 10-letter strings are there? 26^{10}.) If objects are not replaced, it generally results in some sort of factorial. (See above examples.)

Why the names *binomial* and *multinomial coefficients*: First, the binomial. Consider $(x+y)^n$. When computing this product, we take one variable (x or y) from each $(x+y)$ term, and multiply them together. Each such multiplication can be thought of as a string n characters long with only x and y allowed as letters. So for a string n characters long, we choose k of them to be x, leaving the other $(n-k)$ of them to be y. The order we select the positions does not matter because we will re-order them so that all xs are together and all ys are together when we are done. The positions are chosen without replacement, and are all distinguishable from each other. So there are $\binom{n}{k}$ different ways to get a $x^k y^{n-k}$ term, and therefore it has coefficient $\binom{n}{k}$. By similar reasoning, $\binom{n}{a_1,a_2,a_3,\ldots,a_k}$ is the coefficient of the $x_1^{a_1}x_2^{a_2}x_3^{a_3}\ldots x_k^{a_k}$ term of $(x_1+x_2+x_3+\ldots+x_k)^n$.

B.3 More Than One Way to Count

Now even though for any problem there is one right answer, it does not follow that there is only one right way of obtaining that answer.

Example B.3.1. *There are n people. How many different ways are there of forming a k-member committee for $3 \le k \le n$ given that one of them is the chairman of the committee? Bonus question: Why do I set 3 as a lower limit?*

Solution 1: Choose 1 person as chairman, then the rest of the $k-1$ regular members from the $n-1$ people left, to get

$$\binom{n}{1}\binom{n-1}{k-1}.$$

Solution 2: Choose the $k-1$ regular members first, then choose the chairman from the remaining $n-k+1$ people to get

$$\binom{n}{k-1}\binom{n-k+1}{1}.$$

Solution 3: Choose all k committee members, then choose the chairman from among the committee, to get

$$\binom{n}{k}\binom{k}{1}.$$

Solution 4: We can look at this as choosing 1 chairman, $k-1$ regular members, and $n-k$ non-members simultaneously, to get

$$\binom{n}{1,\ k-1,\ n-k}.$$

Now all of these count the same thing, so they should evaluate to the same expression, and they all do:

$$\frac{n!}{(k-1)!(n-k)!}.$$ □

The reason I put this in is to show that it does not matter exactly how you count something, as long as you count all of what you are supposed to (and nothing else). Here's another example.

Example B.3.2. *How many different ways are there to place n identical rooks on an n by n chessboard such that no rook threatens any other rook?*

Solution 1: The first rook can be placed in any of n^2 spaces. Without loss of generality, assume that it is the lower left corner. There are obviously $(n-1)^2$ spaces left for the second rook. Without loss of generality, again assume that we choose the lower left space left available, to obviously leave $(n-2)^2$ spaces left not threatened by either rook, so there are $n^2(n-1)^2$ choices for the first two rooks. This continues until we get $n!^2$. However, each rook is identical, so it does not matter what order we place them down. So divide by $n!$ to obtain simply $n!$.

Solution 2: Number the columns from left to right and the rows from top to bottom. Each row can hold only 1 rook, so there must be 1 per row; same for columns. So take an n-character string with possible values in each spot ranging from 1 to n. The place in the string indicates the row (i.e. first space is for row 1), and the character in that space indicates the column (a '4' in the first space indicates that the rook in row 1 goes in the column 4. This is simply an arrangement without replacement of the numbers from 1 to n, and we know that there are $n!$ of them.

□

B.4 How to Interpret Problems

By now we have gone through elementary counting, and that there can be several different ways to count any given problem. There is still a gap between counting something and knowing what it is to count. Here is a problem that tripped many people up on the first midterm:

Example B.4.1 (Yip Yips). *On the planet of the Yip Yips, the natives play poker with a standard deck of cards, but with 9 cards in hand. Find the probability that a Yip Yip hand will contain the following:*

1. *Three pairs and a triple.*
2. *Two triples and three singletons.*
3. *A flush (including straight flushes).*
4. *A straight (including straight flushes) where the ace can be either high or low.*

Solution:

Because this problem tripped so many people up, I take the answer derivation step by step explicitly so that the *how* of the problem is clear. Immediately we see that we are dealing with card hands, so order does not matter, and objects are chosen without replacement. If we were dealing with the way that the hand is dealt, then order would matter, but objects would still be chosen without replacement. Objects are distinguishable by both rank and suit. So yes they are distinguishable, but we may want to put them in groups later on, depending on what we want to count. Finally, the size of the sample space is the same for all four questions: there are 52 cards, and we choose 9 of them, for a sample space of size $\binom{52}{9}$. All that is left to do is to count the number of events we are interested in and divide by $\binom{52}{9}$ to get the desired answers.

1. *Three pairs and a triple.* What does this mean? It means that there are three different ranks that we have two suits of, and one additional rank that we have three of. We can choose either the pairs or the triple first. Since the difference in difficulty doing it one way or the other is trivial in this circumstance, I choose to count the pairs first, since they are mentioned first. There are 13 ranks, of which we choose 3 for pairs. There are $\binom{13}{3}$ ways this can be done. What does it mean to be a pair? It means that out of 4 suits, we have 2. This can be done $\binom{4}{2}$ ways. Since there are 3 pairs, we do this three times, to obtain $\binom{4}{2}^3$. Now for the triple. There are now 10 available ranks (since we chose 3 of them for the pairs), and we want 1 of them, for a factor of $\binom{10}{1}$.

What does it mean to be a triple? It means that out of 4 suits, we have 3. This can be done $\binom{4}{3}$ ways. So the total number of hands that have 3 pairs and a triple are $\binom{13}{3}\binom{4}{2}^3\binom{10}{1}\binom{4}{3}$. All that is left to do is divide by $\binom{52}{9}$ to get

$$\frac{\binom{13}{3}\binom{4}{2}^3\binom{10}{1}\binom{4}{3}}{\binom{52}{9}}.$$

2. *Two triples and three singletons.* This is rather easy, as we can use some of our reasoning from the first part. In particular, there are still 3 pairs, so we already have $\binom{13}{3}\binom{4}{2}^3$. Now what does it mean to have three singletons? It means that there are 3 separate ranks, each of which we have 1 suit from. So there are 10 available ranks, and we choose 3 of them to get $\binom{10}{3}$, and each rank there are 4 suits and we choose 1 to get $\binom{4}{1}^3$. Multiplying together, and dividing by the sample space size yields

$$\frac{\binom{13}{3}\binom{4}{2}^3\binom{10}{3}\binom{4}{1}^3}{\binom{52}{9}}.$$

3. *A flush (including straight flushes).* What does it mean to be a flush? It means that we have 9 different ranks, all of them of the same suit. So simply choose 9 ranks from 13 to get $\binom{13}{9}$, then choose 1 suit from 4 to get $\binom{4}{1}$. Multiply together, and divide by sample space size to get

$$\frac{\binom{13}{9}\binom{4}{1}}{\binom{52}{9}}.$$

4. *A straight (including straight flushes) where the ace can be either high or low.* What does it mean to be a straight? It means that we have 9 consecutive ranks. Since the ace can be either high or low, we have the following possible values for our straight:

$A, 2, 3, 4, 5, 6, 7, 8, 9$	$2, 3, 4, 5, 6, 7, 8, 9, T$
$3, 4, 5, 6, 7, 8, 9, T, J$	$4, 5, 6, 7, 8, 9, T, J, Q$
$5, 6, 7, 8, 9, T, J, Q, K$	$6, 7, 8, 9, T, J, Q, K, A$

So whatever we come up with for ways of choosing suits, we multiply by 6. Now for each card, there are 4 choices of suit without restriction, and with replacement, so the number of possible suit choices in a straight is 4^9. This leads to a probability of

$$\frac{6 \times 4^9}{\binom{52}{9}}.$$

□

B.5 Challenging Examples

These two challenging problems were given on the first homework assignment, and were considered by many to be the most challenging problems of the 394/395 courses.[2]

[2]They also illustrate the danger for the instructor of assigning a homework problem before deriving the solution!

Example B.5.1. *Tom and Jerry both have standard 52-card decks. They both take the top card and lay it down face up, and repeat this until the decks are exhausted. What is the probability that*

a. *Tom and Jerry match cards exactly every time?*

b. *Tom and Jerry match ranks, but not necessarily suits each time?*

c. *Tom and Jerry match suits, but not necessarily ranks each time?*

d. *Tom and Jerry match ranks, but have different suits each time?*

Solution:

This is a standard 52-card deck. Because all of the cards are used, we can just say that Tom has an arbitrary order, and that Jerry has to have the cards that fulfill the requirements. (The importance of using the entire deck in this problem will become clear next section where we do not use the entire deck.) The other alternative, which is also correct, is to consider the likelihood of Tom's orders as well as Jerry's, but the result of that is simply to multiply both numerator and denominator by 52!. (Aside: multiplying by convenient expressions of 1 or adding convenient expressions of 0 is a very useful technique that we often used to get nice integrals in Math/Stat 395. It is just that this expression of 1 is not convenient in this circumstance.)

a. *Tom and Jerry match cards exactly every time?* Tom has one particular order out of a possible 52!. In order to match, Jerry must have that exact order, for a probability of

$$\frac{1}{52!}.$$

b. *Tom and Jerry match ranks, but not necessarily suits each time?* Tom has one particular order out of a possible 52!. But this time, in order to match, Jerry can mix up suits as long as the ranks match. For instance, if Tom has 2♠ 2♡2♢ 2♣ as the first four cards, Jerry could have 2♡ 2♢ 2♣ 2♠ and still match. So the four cards of each rank (no matter where they are in the order) can be mixed in any way possible. There are 4! such ways for each rank, and 13 ranks, so the probability is

$$\frac{(4!)^{13}}{52!}.$$

c. *Tom and Jerry match suits, but not necessarily ranks each time?* By reversing the role of suit and rank from the previous part, it is easy to see that the probability is

$$\frac{(13!)^4}{52!}.$$

d. *Tom and Jerry match ranks, but have different suits each time?* This one is interesting. It is similar to part b, but instead of allowing all 4! orders, we disallow any order such that Jerry and Tom match any suit. Ultimately, just enumerate the possible orders for the suits of each rank. If Tom has the order ♠ ♡ ♢ ♣, then Jerry might have:

♡♠♣♢	♡♣♠♢	♡♢♣♠
♢♠♣♡	♢♣♠♡	♢♣♡♠
♣♠♡♢	♣♢♠♡	♣♢♡♠

So Jerry has 9 possible orders of suits for each rank, and there are 13 ranks, so the probability is

$$\frac{9^{13}}{52!}.$$

□

Example B.5.2. *Tom and Jerry both have standard 52-card decks. They both take the top card and lay it down face up, and repeat this only three times. What is the probability that*

a. *Tom and Jerry match cards exactly every time?*

b. *Tom and Jerry match ranks, but not necessarily suits each time?*

c. *Tom and Jerry match suits, but not necessarily ranks each time?*

d. *Tom and Jerry match ranks, but have different suits each time?*

Solution:

The first thought might be that with only 3 cards instead of the entire deck, this would be easier than problem 17. This is incorrect because with problem 17 the entire deck was used in each order, eliminating the need for cases and allowing us to focus only on Jerry's order rather than both decks' orders. The second thought might be that it is only a little bit harder than problem 17. This also turned out to be incorrect, at least for all of us in that class (including the grader and instructor).

a. *Tom and Jerry match cards exactly every time?* This is the easiest of the bunch, and because of that, sets up the expectation that the rest will be easy. Tom's first three cards have an order. Jerry has a $\frac{1}{52}$ chance of matching the first card. Given that, he has a $\frac{1}{51}$ chance of matching the second card. Given both of those, he has a $\frac{1}{50}$ chance of matching the third card, for a probability of $\frac{1}{{}_{52}\mathrm{P}_3}$, or

$$\frac{49!}{52!}.$$

b. *Tom and Jerry match ranks, but not necessarily suits each time?* Here is where the difficulty truly begins, because in this situation, Tom might have 1, 2, or 3 different ranks, each with a different probability. And depending on the number of ranks that Tom has, Jerry has a different probability in matching him. Fortunately, since the probabilities of having 1, 2, or 3 different ranks is disjoint, we can add them at the end.

   So the probability that Tom has 1 rank is $\frac{\binom{13}{1}{}_4\mathrm{P}_3}{{}_{52}\mathrm{P}_3}$. That is, 13 possible ranks, 4 different ways to choose the first card, 3 for the second, and 2 for the third, out of $52 \times 51 \times 50$ possible orders. If Tom has 1 rank, then the probability that Jerry matches that rank but not necessarily the suit each time is $\frac{4}{52} \times \frac{3}{51} \times \frac{2}{50}$. So the total probability that Jerry matches Tom with only 1 rank is $\frac{\binom{13}{1}{}_4\mathrm{P}_3^2}{{}_{52}\mathrm{P}_3^2}$.

   Now for 2 ranks. One is a pair, and one is a singleton. There are 13 choices for the pair, and then 12 for the singleton. In the pair, there are 4 choices for the first card and 3 for the second. There are four choices for the singleton, which can appear in any of the three positions, so the probability that Tom has 2 ranks is $\frac{{}_{13}\mathrm{P}_2 \times {}_4\mathrm{P}_2 \times {}_4\mathrm{P}_1 \times 3}{{}_{52}\mathrm{P}_3}$. If Tom has 2 ranks, then the probability that Jerry matches it is $\frac{{}_4\mathrm{P}_2 \times {}_4\mathrm{P}_1}{{}_{52}\mathrm{P}_3}$, where there are 4 choices for the first card of the pair, 3 choices for the second, and 4 choices for the singleton, out of a possible $52 \times 51 \times 50$ orders, for a total of $\frac{{}_{13}\mathrm{P}_2 \times {}_4\mathrm{P}_2^2 \times {}_4\mathrm{P}_1^2 \times 3}{{}_{52}\mathrm{P}_3^2}$.

   Next is the 3 rank case. This is easier than the 2 rank case, actually. For Tom: 13 choices of rank for the first card, 12 for the second, 11 for the third, and for each one 4 choices of suit to obtain $\frac{{}_{13}\mathrm{P}_3 \times 4^3}{{}_{52}\mathrm{P}_3}$. For Jerry, there are 4 choices for the first card (any suit of the given rank), 4 more for the second, and 4 more for the third, for a total of $\frac{4^3}{{}_{52}\mathrm{P}_3}$. So the total probability of Jerry matching Tom with 3 ranks is $\frac{{}_{13}\mathrm{P}_3 \times 4^6}{{}_{52}\mathrm{P}_3^2}$.

   Finally, add all the probabilities together to obtain:

$$\frac{\binom{13}{1}{}_4\mathrm{P}_3^2 + \left({}_{13}\mathrm{P}_2 \times {}_4\mathrm{P}_2^2 \times {}_4\mathrm{P}_1^2 \times 3\right) + \left({}_{13}\mathrm{P}_3 \times 4^6\right)}{{}_{52}\mathrm{P}_3^2}.$$

c. *Tom and Jerry match suits, but not necessarily ranks each time?* Now we deal with matching suits instead of ranks, and we proceed in a similar fashion to part b.

The probability that Tom has 1 suit is $\frac{\binom{4}{1}{}_{13}\mathrm{P}_3}{{}_{52}\mathrm{P}_3}$. That is 4 possible suits, 13 ways to choose the first card, 12 for the second, and 11 for the third, out of $52 \times 51 \times 50$ possible orders. If Tom has 1 suit, then the probability that Jerry matches that suit but not necessarily the rank each time is $\frac{13}{52} \times \frac{12}{51} \times \frac{11}{50}$. So the total probability that Jerry matches Tom with only 1 suit is $\frac{\binom{4}{1}{}_{13}\mathrm{P}_3^2}{{}_{52}\mathrm{P}_3^2}$.

Now for 2 suits. One is a doubleton (has only two cards) and the other is a singleton. There are 4 choices for the doubleton suit, and then 3 for the singleton. In the doubleton, there are 13 choices for the first card and 12 for the second. There are 13 choices for the singleton, which can appear in any of the 3 positions. So the probability that Tom has 2 suits is $\frac{{}_4\mathrm{P}_2 \cdot {}_{13}\mathrm{P}_2 \cdot {}_{13}\mathrm{P}_1 \cdot 3}{{}_{52}\mathrm{P}_3}$. If Tom has 2 suits, then the probability that Jerry matches it is $\frac{{}_{13}\mathrm{P}_2 {}_{13}\mathrm{P}_1}{{}_{52}\mathrm{P}_3}$, where there are 13 choices for the first card of the doubleton, 12 choices for the second, and 13 choices for the singleton, out of a possible $52 \cdot 51 \cdot 50$ orders. Multiplying the probabilities together, the probability that Jerry matches Tom with 2 suits is $\frac{{}_4\mathrm{P}_2 \cdot {}_{13}\mathrm{P}_2^2 \cdot {}_{13}\mathrm{P}_1^2 \cdot 3}{({}_{52}\mathrm{P}_3)^2} = \frac{13^4 \cdot 3^3 \cdot 2^6}{(52 \cdot 51 \cdot 50)^2}$.

Next is the 3 suit case. As in part b, this is easier than the 2 suit case. For Tom: 4 choices of suit for the first card, 3 for the second, 2 for the third, and for each suit 13 choices of rank to obtain $\frac{{}_4\mathrm{P}_3 \cdot \cdot 13^3}{{}_{52}\mathrm{P}_3}$. For Jerry, there are 13 choices for the first card (any rank of the given suit), 13 more for the second, and 13 more for the third, for a total of $\frac{13^3}{{}_{52}\mathrm{P}_3 \cdot 51 \cdot 50}$. So the total probability of Jerry matching Tom with 3 suits is $\frac{{}_4\mathrm{P}_3 \cdot 13^6}{{}_{52}\mathrm{P}_3^2} = \frac{4 \cdot 3 \cdot 2 \cdot 13^6}{(52 \cdot 51 \cdot 50)^2}$.

Finally, add all the probabilities together to obtain

$$\frac{\binom{4}{1}{}_{13}\mathrm{P}_3^2 + {}_4\mathrm{P}_2 \cdot {}_{13}\mathrm{P}_2^2 \cdot {}_{13}\mathrm{P}_1^2 \cdot 3 + {}_4\mathrm{P}_3 \cdot 13^6}{{}_{52}\mathrm{P}_3^2} = \frac{\left(13^2 \cdot 11^2 \cdot 3^2 \cdot 2^6\right) + \left(13^4 \cdot 3^3 \cdot 2^6\right) + \left(13^6 \cdot 3 \cdot 2^3\right)}{(52 \cdot 51 \cdot 50)^2}.$$

d. *Tom and Jerry match ranks, but have different suits each time?* This is going to be trickier than the rest of the questions, since not only do we have to find out how many ways there are to match ranks but avoid suits, but we have to do it for ranks with 1, 2, and 3 cards in them. Unfortunately, the easiest way is still by brute enumeration.

If a rank has only 1 suit, say ♠, then there are 3 ways to avoid matching it: ♡, ♢ and ♣.

If a rank has 2 suits, say (♠, ♡), then there are 7 ways to avoid matching it: (♡, ♠), (♡, ♢), (♡, ♣), (♢, ♠), (♢, ♣), (♣, ♠), and (♣, ♢).

If a rank has 3 suits, say (♠, ♡, ♢), then there are 11 ways to avoid matching it: (♡, ♠, ♣), (♡, ♢, ♠), (♡, ♢, ♣), (♡, ♣, ♠), (♢, ♠, ♡), (♢, ♠, ♣), (♢, ♣, ♠), (♢, ♣, ♡), (♣, ♠, ♡), (♣, ♢, ♠), and (♣, ♢, ♡).

Now that we have that information, we look at what Tom actually has.

From part b, the probability that Tom has 1 rank is $\frac{\binom{13}{1}{}_4\mathrm{P}_3}{{}_{52}\mathrm{P}_3}$. From above, we see that there are 11 ways that Jerry can match rank but not suit out of ${}_{52}\mathrm{P}_3$ possible ways, for a a total probability of $\frac{\binom{13}{1}{}_4\mathrm{P}_3 \cdot 11}{{}_{52}\mathrm{P}_3^2}$.

Again from part b, the probability that Tom has 2 ranks is $\frac{{}_4\mathrm{P}_2 \cdot {}_{13}\mathrm{P}_2 \cdot {}_{13}\mathrm{P}_1 \cdot 3}{{}_{52}\mathrm{P}_3}$. One of those ranks is a pair, and so there are 7 ways to match rank but not suit. The other rank is a singleton, and there are 3 ways to match that rank but not suit. So there are 21 matches out of ${}_{52}\mathrm{P}_3$ ways, for a total probability of $\frac{{}_4\mathrm{P}_2 \cdot {}_{13}\mathrm{P}_2 \cdot {}_{13}\mathrm{P}_1 \cdot 63}{{}_{52}\mathrm{P}_3^2}$.

Once more, from part b, the probability that Tom has 3 ranks is $\frac{{}_4\mathrm{P}_3 \cdot 13^3}{{}_{52}\mathrm{P}_3}$. With each rank, there are 3 valid matches, so there are 3^3 total valid matches out of ${}_{52}\mathrm{P}_3$ orders, for a probability of $\frac{{}_4\mathrm{P}_3 \cdot 3^3 \cdot 13^3}{{}_{52}\mathrm{P}_3^2}$.

Summing these all together obtains the total probability of

$$\frac{\left(\binom{13}{1} \cdot {}_4\mathrm{P}_3 \cdot 11\right) + \left({}_4\mathrm{P}_2 \cdot {}_{13}\mathrm{P}_2 \cdot {}_{13}\mathrm{P}_1 \cdot 63\right) + \left({}_4\mathrm{P}_3 \cdot 3^3 \cdot 13^3\right)}{{}_{52}\mathrm{P}_3^2}.$$ □

C Table of Common Distributions

C.1 Discrete Distributions

C.1.1 Binomial

Denoted $X \sim \mathrm{B}(n, p)$.

$$p_X(k) = \binom{n}{k} p^k q^{n-k}, \quad n \in \mathbb{Z}^+, \;\; k \in \{0, 1, 2, \ldots, n\}, \;\; p \in [0, 1], \;\; q = 1 - p.$$

$$\mathrm{E}[X] = np, \quad \mathrm{Var}(X) = npq, \quad M_X(t) = \left((1-p) + pe^t\right)^n.$$

This is used in situations where we have n trials, and each trial just has two choices (0 or 1, success or failure, heads or tails, etc.), where 1 happens with probability p and 0 has probability q. X is the number of 1's (successes, heads, etc) in those n trials.

If $n = 1$, this is called a *Bernoulli* distribution.

C.1.2 Geometric

Denoted $X \sim \mathrm{Geometric}(p)$.

$$p_X(k) = pq^{k-1}, \quad k \in \mathbb{Z}^+, \quad p \in [0, 1], \;\; q = 1 - p.$$

$$\mathrm{E}[X] = \frac{1}{p}, \quad \mathrm{Var}(X) = \frac{1-p}{p^2}, \quad M_X(t) = \frac{pe^t}{1-(1-p)e^t}.$$

This is used in situations where we have n trials, and each trial just has two choices (0 or 1, success or failure, heads or tails, etc.), where 1 happens with probability p and 0 has probability q. X is the number of trials needed until we have 1 success.

C.1.3 Hypergeometric

No special notation for this one.

$$p_X(k) = \frac{\binom{M}{k}\binom{N-M}{K-k}}{\binom{N}{K}}, \quad \max(0, M - (N - K)) \le k \le M; \quad M, N, K \in \mathbb{Z}^*.$$

$$\mathrm{E}[X] = \frac{KM}{N}, \quad \mathrm{Var}(X) = \frac{KM}{N}\frac{(N-M)(N-K)}{N(N-1)}, \quad M_X(t) = \text{Something really ugly.}$$

This is used when you are choosing K items out of N of them (each equally likely to be chosen) and you want to know the probability that k out of K of them are of one type (with a population of M).

C.1.4 Negative Binomial

No special notation for this one.

$$p_X(k) = \binom{k-1}{r-1} p^r q^{k-r}, \quad k, r \in \mathbb{Z}^+; \quad k \ge r; \quad p \in [0, 1], \quad q = 1 - p.$$

$$\mathrm{E}[X] = \frac{r}{p}, \quad \mathrm{Var}(X) = \frac{r(1-p)}{p^2}, \quad M_X(t) = \left(\frac{pe^t}{1-(1-p)e^t}\right)^r.$$

This is used in situations where we have n trials, and each trial just has two choices (0 or 1, success or failure, heads or tails, etc.), where 1 happens with probability p and 0 has probability q. X is the number of trials needed until we have r successes.

C.1.5 Poisson

Denoted $X \sim \mathrm{P}(\lambda^*)$.

$$p_X(k) = e^{-\lambda^*}\frac{(\lambda^*)^k}{k!}, \quad k \in \mathbb{Z}^*,\ \lambda^* > 0.$$

$$\mathrm{E}[X] = \lambda^*, \quad \mathrm{Var}(X) = \lambda^*, \quad M_X(t) = e^{\lambda(e^t-1)}.$$

λ^* is unitless. This is used in a *Poisson process* over time (or some other quantity) t, where

- Instead of X, we use $N(t)$, which describes the number of (rare) events that happened up to time t.
- $\lambda^* = \lambda t$, where λ is a rate of some event per quantity in some unit, and t is that time measured in the same unit.

C.2 Continuous Distributions

C.2.1 Beta

Denoted $X \sim \mathrm{Beta}(a, b)$.

$$f_X(x) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1; \quad B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}; \quad a, b > 0$$

$$\mathrm{E}[X] = \frac{a}{a+b}, \quad \mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}, \quad M_X(t) = 1 + \sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{t^k}{k!}.$$

C.2.2 Cauchy

No special notation for this one.

$$f_X(x) = \frac{1}{\pi(1+(x-\theta)^2)} \quad x, \theta \in \mathbb{R}; \quad \alpha, \beta > 0.$$

$\mathrm{E}[X], \mathrm{Var}(X)$ and $M_X(t)$ do not exist (if you try to calculate them, you will get ∞).

If Y_1 and Y_2 are iid $N(0,1)$ random variables, then $X = Y_1/Y_2$ has a Cauchy distribution with $\theta = 0$.

C.2.3 Chi-Squared

Denoted $X \sim \chi^2_n$.

$$f_X(x) = \frac{x^{n/2-1}e^{-x/2}}{\Gamma\left(\frac{n}{2}\right)2^{n/2}}$$

$$\mathrm{E}[X] = n, \quad \mathrm{Var}(X) = 2n, \quad M_X(t) = \left(\frac{1}{1-2t}\right)^{n/2}.$$

This is the same as a Gamma$\left(\frac{n}{2}, \frac{1}{2}\right)$ distribution. If $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathrm{N}(0,1)$, then $\sum_{i=1}^n X_i^2 \sim \chi^2_n$.

C.2.4 Exponential

Denoted $X \sim \mathrm{Exp}(\lambda)$.

$$f_X(x) = \lambda e^{-\lambda x} \quad x \geq 0, \quad \lambda > 0.$$

$$\mathrm{E}[X] = \frac{1}{\lambda}, \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}, \quad M_X(t) = \frac{\lambda}{\lambda - t}.$$

In a Poisson process with rate λ, the inter-arrival times are exponentially distributed with parameter λ.

C.2.5 Gamma

Denoted $X \sim \mathrm{Gamma}(\alpha, \beta)$.

$$f_X(x) = \frac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)} \quad x \geq 0; \quad \alpha, \beta > 0.$$

$$\mathrm{E}[X] = \frac{\alpha}{\beta}, \quad \mathrm{Var}(X) = \frac{\alpha}{\beta^2}, \quad M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^\alpha.$$

If X is the sum of n iid $\mathrm{Exp}(\lambda)$ random variables, then $X \sim \mathrm{Gamma}(n, \lambda)$.

C.2.6 Normal

Denoted $X \sim \mathrm{N}(\mu, \sigma)$ or $X \sim \mathrm{N}(\mu, \sigma^2)$. To avoid confusion, we write something like "$X \sim N(a, \sigma^2 = b)$".

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}, \quad x \in (-\infty, \infty), \quad \mu \in (-\infty, \infty), \quad \sigma \in (0, \infty).$$

$$\mathrm{E}[X] = \mu, \quad \mathrm{Var}(X) = \sigma^2, \quad M_X(t) = e^{\mu t + \sigma^2 t^2/2}.$$

This is also called the *Gaussian* distribution, or more informally, the bell curve. This is the most important distribution in statistics because of the *Central Limit Theorem*, which states that if $X_1, X_2, \ldots$ are identically and independently distributed for *any* distribution, the sample mean minus the expected value, divided by the standard deviation, is distributed as $\mathrm{N}(0,1)$.

Note that if $X \sim N(\mu, \sigma)$, then $\frac{X-\mu}{\sigma} \sim N(0,1)$, which is called the *standard normal distribution*, and often referred to as Z.

Since an integral of this distribution is difficult, values are tabulated on tables. Values are given for $\Phi(z) = \mathrm{P}(Z \leq z)$ for $z \geq 0$. Note that, since the distribution is symmetric,

$$\Phi(-z) = \mathrm{P}(Z \leq -z) = 1 - \mathrm{P}(Z \leq z) = 1 - \Phi(z).$$

C.2.7 Uniform

Denoted $X \sim \mathrm{U}(\alpha, \beta)$.

$$f_X(x) = \frac{1}{\beta - \alpha}, \quad \alpha < x < \beta$$

$$\mathrm{E}[X] = \frac{\beta + \alpha}{2}, \quad \mathrm{Var}(X) = \frac{(\beta - \alpha)^2}{12}, \quad M_X(t) = \frac{e^{bt} - e^{at}}{(b-a)t}.$$

This is arguably the simplest continuous distribution. It is often used (especially in Bayesian statistics) as a starting point when we have no idea what the distribution looks like, other than the fact that the values will range from α to β.

Bibliography

Birnbaum, Z. W. (1962), *Introduction to Probability and Mathematical Statistics*, Harper and Brothers, New York.

Bóna, M. (2006), *A Walk through Combinatorics: An Introduction to Enumeration and Graph Theory*, second edn, World Scientific Publishing Co., Singapore.

D'Angelo, J. P. and West, D. B. (2000), *Mathematical Thinking: Problem-Solving and Proofs*, second edn, Prentice-Hall, Inc., Upper Saddle River, NJ.

Feller, W. (1968), *An Introduction to Probability Theory and Its Applications, Volume 1*, John Wiley and Sons, Inc., New York.

Ferguson, T. S. (1996), *A Course in Large Sample Theory*, Chapman and Hall/CRC, Boca Raton, FL.

Parzen, E. (1960), *Modern Probability Theory and Its Applications*, John Wiley and Sons, Inc., New York.

Ross, S. (2002), *A First Course in Probability*, sixth edn, Pearson Prentice Hall, Upper Saddle River, NJ.

Ross, S. (2006), *A First Course in Probability*, seventh edn, Pearson Prentice Hall, Upper Saddle River, NJ.

Skiena, S. (2001), *Calculated Bets*, Cambridge University Press, Cambridge, UK.