STAT/MATH 511

PROBABILITY

Fall, 2007

Lecture Notes

Joshua M. Tebbs

Department of Statistics

University of South Carolina

Contents

1 Probability

Complementary reading: Chapter 2 (WMS).

1.1 Introduction

TERMINOLOGY: The text defines **probability** as a measure of one's belief in the occurrence of a future event. It is also sometimes called "the mathematics of uncertainty."

EVENTS: Here are some events we may wish to assign probabilities to:

- tomorrow's temperature exceeding 80 degrees
- manufacturing a defective part
- concluding one fertilizer is superior to another when it isn't
- the NASDAQ losing 5 percent of its value
- you earning a "B" or better in this course.

ASSIGNING PROBABILITIES TO EVENTS: How do we **assign** probabilities to events? There are three general approaches.

1. *Subjective approach.*
   - this is based on feeling and may not even be scientific.
2. *Relative frequency approach.*
   - this approach can be used when some **random phenomenon** is observed repeatedly under identical conditions.
3. *Axiomatic approach.* This is the approach we will take in this course.

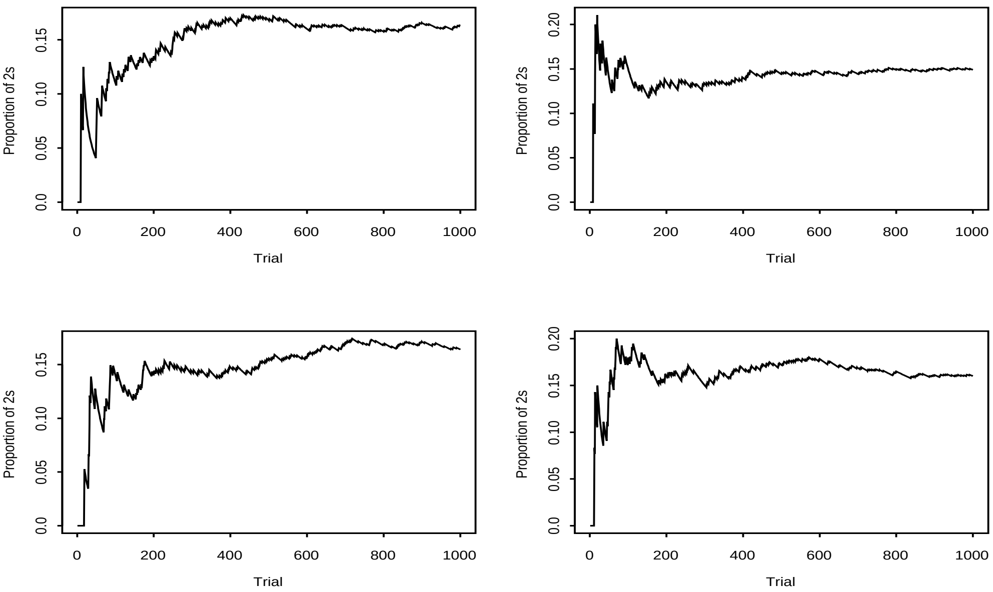

Figure 1.1: *The proportion of tosses which result in a "2"; each plot represents* 1,000 *rolls of a fair die.*

Example 1.1. *An example illustrating the relative frequency approach to probability.* Suppose we roll a die 1000 times and record the number of times we observe a "2." Let A denote this event. The **relative frequency approach** says that

$$P(A) \approx \frac{\text{number of times } A \text{ occurs}}{\text{number of trials performed}} = \frac{n(A)}{n},$$

where $n(A)$ denotes the **frequency** of the event, and n denotes the number of trials performed. The ratio $n(A)/n$ is sometimes called the **relative frequency**. The symbol $P(A)$ is shorthand for "the probability that A occurs."

RELATIVE FREQUENCY APPROACH: Continuing with our example, suppose that $n(A) = 158$. Then, we would **estimate** $P(A)$ with $158/1000 = 0.158$. If we performed this experiment repeatedly, the relative frequency approach says that

$$n(A)/n \longrightarrow P(A),$$

as $n \to \infty$. Of course, if the die is **unbiased**, $n(A)/n \to P(A) = 1/6$. □

1.2 Sample spaces

TERMINOLOGY: In probability applications, it is common to perform some **random experiment** and then observe an outcome. The set of all possible outcomes for an experiment is called the **sample space**, hereafter denoted by S.

Example 1.2. The Michigan state lottery calls for a three-digit integer to be selected:

$$S = \{000, 001, 002, ..., 998, 999\}. \ \Box$$

Example 1.3. An industrial experiment consists of observing the lifetime of a certain battery. If lifetimes are measured in hours, the sample space could be any one of

$$S_1 = \{w : w \geq 0\} \quad S_2 = \{0, 1, 2, 3, ..., \} \quad S_3 = \{\text{defective, not defective}\}. \ \Box$$

MORAL: Sample spaces are **not** unique; in fact, how we define the sample space has a direct influence on how we assign probabilities to events.

1.3 Basic set theory

TERMINOLOGY: A **countable set** A is one whose elements can be put into a one-to-one correspondence with $\mathcal{N} = \{1, 2, ..., \}$, the set of natural numbers (i.e., there exists an injection with domain A and range $\mathcal{N}$). A set that is not countable is called an **uncountable set**.

TERMINOLOGY: Countable sets can be further divided up into two types. A **countably infinite set** has an infinite number of elements. A **countably finite set** has a finite number of elements.

TERMINOLOGY: Suppose that S is a nonempty set. We say that A is a **subset** of S, and write $A \subset S$ (or $A \subseteq S$), if

$$\omega \in A \Rightarrow \omega \in S.$$

In probability applications, S will denote a **sample space**, A will represent an **event** to which we wish to assign a probability, and ω usually denotes a possible **experimental outcome**. If $\omega \in A$, we would say that "the event A has occurred."

TERMINOLOGY: The **null set**, denoted as $\emptyset$, is the set that contains no elements.

TERMINOLOGY: The **union** of two sets is the set of all elements in either set or both. We denote the union of two sets A and B as $A \cup B$. In ω notation,

$$A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B\}.$$

TERMINOLOGY: The **intersection** of two sets A and B is the set containing those elements which are in both sets. We denote the intersection of two sets A and B as $A \cap B$. In ω notation,

$$A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}.$$

EXTENSION: We can extend the notion of unions and intersections to more than two sets. Suppose that $A_1, A_2, ..., A_n$ is a **finite** sequence of sets. The union of these n sets is

$$\bigcup_{j=1}^{n} A_j = A_1 \cup A_2 \cup \cdots \cup A_n = \{\omega : \omega \in A_j \text{ for at least one } j\},$$

and the intersection of the n sets is

$$\bigcap_{j=1}^{n} A_j = A_1 \cap A_2 \cap \cdots \cap A_n = \{\omega : \omega \in A_j \text{ for all } j\}.$$

EXTENSION: Suppose that $A_1, A_2, ...$ is a **countable** sequence of sets. The union and intersection of this infinite collection of sets is

$$\begin{aligned} \bigcup_{j=1}^{\infty} A_j &= \{\omega : \omega \in A_j \text{ for at least one } j\} \\ \bigcap_{j=1}^{\infty} A_j &= \{\omega : \omega \in A_j \text{ for all } j\}. \end{aligned}$$

Example 1.4. Define the sequence of sets $A_j = [1, 1 + 1/j)$, for $j = 1, 2, ...,$. Then,

$$\bigcup_{j=1}^{\infty} A_j = [1, 2) \quad \text{and} \quad \bigcap_{j=1}^{\infty} A_j = \{1\}. \ \square$$

TERMINOLOGY: The **complement** of a set A is the set of all elements not in A (but still in S). We denote the complement as $\overline{A}$. In ω notation,

$$\overline{A} = \{\omega \in S : \omega \notin A\}$$

TERMINOLOGY: We say that A is a **subset** of B, and write $A \subset B$ (or $A \subseteq B$) if $\omega \in A \Rightarrow \omega \in B$. Thus, if A and B are events in an experiment and $A \subset B$, then, if A occurs, B must occur as well.

Distributive Laws:

1. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
2. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

DeMorgans Laws:

1. $\overline{A \cap B} = \overline{A} \cup \overline{B}$
2. $\overline{A \cup B} = \overline{A} \cap \overline{B}$

TERMINOLOGY: We call two events A and B **mutually exclusive**, or **disjoint**, if $A \cap B = \emptyset$. Extending this definition to a finite or countable collection of sets is obvious.

1.4 Properties of probability

THE THREE AXIOMS OF PROBABILITY: Given a nonempty sample space S, the measure $P(A)$ is a **set function** satisfying three axioms:

(1) $P(A) \geq 0$, for every $A \subseteq S$

(2) $P(S) = 1$

(3) If $A_1, A_2, ...$ is a countable sequence of **pairwise mutually exclusive** events (i.e., $A_i \cap A_j = \emptyset$, for $i \neq j$) in S, then

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

IMPORTANT RESULTS: The following results are important properties of the **probability set function** $P(\cdot)$, and each follows from the Kolmolgorov Axioms (those just stated). All events below are assumed to be subsets of S.

1. **Complement rule:** For any event A,

$$P(A) = 1 - P(\overline{A}).$$

Proof. Note that $S = A \cup \overline{A}$. Thus, since A and $\overline{A}$ are disjoint, $P(A \cup \overline{A}) = P(A) + P(\overline{A})$ by Axiom 3. By Axiom 2, $P(S) = 1$. Thus,

$$1 = P(S) = P(A \cup \overline{A}) = P(A) + P(\overline{A}). \ \Box$$

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

Proof. Take $A = \emptyset$ and $\overline{A} = S$. Use the last result and Axiom 2. □

3. **Monotonicity property:** Suppose that A and B are two events such that $A \subset B$. Then, $P(A) \leq P(B)$.

Proof. Write $B = A \cup (B \cap \overline{A})$. Clearly, A and $(B \cap \overline{A})$ are disjoint. Thus, by Axiom 3, $P(B) = P(A) + P(B \cap \overline{A})$. Because $P(B \cap \overline{A}) \geq 0$, we are done. □

4. For any event A, $P(A) \leq 1$.

Proof. Since $A \subset S$, this follows from the monotonicity property and Axiom 2. □

5. **Inclusion-exclusion:** Suppose that A and B are two events. Then,

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

Proof. Write $A \cup B = A \cup (\overline{A} \cap B)$. Then, since A and $(\overline{A} \cap B)$ are disjoint, by Axiom 3,

$$P(A \cup B) = P(A) + P(\overline{A} \cap B).$$

Now, write $B = (A \cap B) \cup (\overline{A} \cap B)$. Clearly, $(A \cap B)$ and $(\overline{A} \cap B)$ are disjoint. Thus, again, by Axiom 3,

$$P(B) = P(A \cap B) + P(\overline{A} \cap B).$$

Combining the last two statements gives the result. □

Example 1.5. The probability that train 1 is on time is 0.95, and the probability that train 2 is on time is 0.93. The probability that both are on time is 0.90.

(a) What is the probability that **at least one** train is on time?
SOLUTION: Denote by A_i the event that train i is on time (for $i = 1, 2$). Then,

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) = 0.95 + 0.93 - 0.90 = 0.98. \ \Box$$

(b) What is the probability that **neither** train is on time?
SOLUTION: By DeMorgan's Law

$$P(\overline{A}_1 \cap \overline{A}_2) = P(\overline{A_1 \cup A_2}) = 1 - P(A_1 \cup A_2) = 1 - 0.98 = 0.02. \ \Box$$

EXTENSION: The **inclusion-exclusion** formula can be extended to any finite sequence of sets $A_1, A_2, ..., A_n$. For example, if $n = 3$,

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) \\ &\quad - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3). \end{aligned}$$

In general, the inclusion-exclusion formula can be written for any finite sequence:

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_j\right) &= \sum_{i=1}^{n} P(A_i) - \sum_{i_1 < i_2} P(A_{i_1} \cap A_{i_2}) + \sum_{i_1 < i_2 < i_3} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \\ &\quad \cdots + (-1)^{n+1} P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n}). \end{aligned}$$

Of course, if the sets $A_1, A_2, ..., A_n$ are disjoint, then we arrive back at

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

a result implied by Axiom 3.

1.5 Discrete probability models and events

TERMINOLOGY: If a sample space for an experiment contains a finite or countable number of sample points, we call it a **discrete sample space**.

- **Finite:** "number of sample points $< \infty$."
- **Countable:** "number of sample points may equal ∞, but can be counted; i.e., sample points may be put into a 1:1 correspondence with $\mathcal{N} = \{1, 2, ..., \}$."

Example 1.6. A standard roulette wheel contains an array of numbered compartments referred to as "pockets." The pockets are either red, black, or green. The numbers 1 through 36 are evenly split between red and black, while 0 and 00 are green pockets. On the next play, one may be interested in the following events:

$$\begin{aligned} A_1 &= \{13\} \\ A_2 &= \{\text{red}\} \\ A_3 &= \{0, 00\} \end{aligned}$$

TERMINOLOGY: A **simple event** is one that can not be decomposed. That is, a simple event corresponds to exactly one sample point ω. **Compound events** are those events that contain more than one sample point. In Example 1.6, because A_1 only contains one sample point, it is a simple event. The events A_2 and A_3 contain more than one sample point; thus, they are compound events.

STRATEGY: Computing the probability of a **compound** event can be done by

(1) identifying all sample points associated with the event

(2) adding up the probabilities associated with each sample point.

NOTATION: We have used ω to denote an element in a set (i.e., a sample point in an event). In a more probabilistic spirit, your authors use the symbol E_i to denote the ith sample point (i.e., simple event). Thus, if A denotes any **compound event**,

$$P(A) = \sum_{i:E_i \in A} P(E_i).$$

We simply sum up the simple event probabilities $P(E_i)$ for all i such that $E_i \in A$.

RESULT: Suppose a discrete sample space S contains $N < \infty$ sample points, each of which are **equally likely**. If the event A consists of n_a sample points, then $P(A) = n_a/N$.

Proof. Write $S = E_1 \cup E_2 \cup \cdots \cup E_N$, where E_i corresponds to the ith sample point; $i = 1, 2, ..., N$. Then,

$$1 = P(S) = P(E_1 \cup E_2 \cup \cdots \cup E_N) = \sum_{i=1}^{N} P(E_i).$$

Now, as $P(E_1) = P(E_2) = \cdots = P(E_N)$, we have that

$$1 = \sum_{i=1}^{N} P(E_i) = NP(E_1),$$

and, thus, $P(E_1) = \frac{1}{N} = P(E_2) = \cdots = P(E_N)$. Without loss of generality, take $A = E_1 \cup E_2 \cup \cdots \cup E_{n_a}$. Then,

$$P(A) = P(E_1 \cup E_2 \cup \cdots \cup E_{n_a}) = \sum_{i=1}^{n_a} P(E_i) = \sum_{i=1}^{n_a} \frac{1}{N} = n_a/N. \quad \square$$

1.6 Tools for counting sample points

1.6.1 The multiplication rule

MULTIPLICATION RULE: Consider an experiment consisting of $k \geq 2$ "stages," where

$$\begin{aligned} n_1 &= \text{number of ways stage 1 can occur} \\ n_2 &= \text{number of ways stage 2 can occur} \\ &\vdots \\ n_k &= \text{number of ways stage } k \text{ can occur} \end{aligned}$$

Then, there are

$$\prod_{i=1}^{k} n_i = n_1 \times n_2 \times \cdots \times n_k$$

different outcomes in the experiment.

Example 1.7. An experiment consists of rolling two dice. Envision stage 1 as rolling the first and stage 2 as rolling the second. Here, $n_1 = 6$ and $n_2 = 6$. By the multiplication rule, there are $n_1 \times n_2 = 6 \times 6 = 36$ different outcomes. $\square$

Example 1.8. In a field experiment, I want to form all possible treatment combinations among the three factors:

Factor 1: Fertilizer (60 kg, 80 kg, 100kg: 3 levels)

Factor 2: Insects (infected/not infected: 2 levels)

Factor 3: Temperature (70F, 90F: 2 levels).

Here, $n_1 = 3, n_2 = 2$, and $n_3 = 2$. Thus, by the multiplication rule, there are $n_1 \times n_2 \times n_3 = 12$ different treatment combinations. □

Example 1.9. Suppose that an Iowa license plate consists of seven places; the first three are occupied by letters; the remaining four with numbers. Compute the total number of possible orderings if

(a) there are no letter/number restrictions.

(b) repetition of letters is prohibited.

(c) repetition of numbers is prohibited.

(d) repetitions of numbers and letters are prohibited.

Answers:

(a) $26 \times 26 \times 26 \times 10 \times 10 \times 10 \times 10 = 175,760,000$

(b) $26 \times 25 \times 24 \times 10 \times 10 \times 10 \times 10 = 156,000,000$

(c) $26 \times 26 \times 26 \times 10 \times 9 \times 8 \times 7 = 88,583,040$

(d) $26 \times 25 \times 24 \times 10 \times 9 \times 8 \times 7 = 78,624,000$

1.6.2 Permutations

TERMINOLOGY: A **permutation** is an arrangement of distinct objects in a particular order. *Order is important.*

PROBLEM: Suppose that we have n distinct objects and we want to **order** (or **permute**) these objects. Thinking of n slots, we will put one object in each slot. There are

- n different ways to choose the object for slot 1,
- $n-1$ different ways to choose the object for slot 2,
- $n-2$ different ways to choose the object for slot 3,

and so on, down to

- 2 different ways to choose the object for slot $(n-1)$, and
- 1 way to choose for the last slot.

PUNCHLINE: By the multiplication rule, there are $n(n-1)(n-2)\cdots(2)(1) = n!$ different ways to order (permute) the n distinct objects.

Example 1.10. My bookshelf has 10 books on it. How many ways can I permute the 10 books on the shelf? ANSWER: $10! = 3,628,800$. □

Example 1.11. Now, suppose that in Example 1.10 there are 4 math books, 2 chemistry books, 3 physics books, and 1 statistics book. I want to order the 10 books so that all books of the same subject are together. How many ways can I do this?
SOLUTION: Use the multiplication rule.

Stage 1	Permute the 4 math books	4!
Stage 2	Permute the 2 chemistry books	2!
Stage 3	Permute the 3 physics books	3!
Stage 4	Permute the 1 statistics book	1!
Stage 5	Permute the 4 subjects $\{m, c, p, s\}$	4!

Thus, there are $4! \times 2! \times 3! \times 1! \times 4! = 6912$ different orderings. □

PERMUTATIONS: With a collection of n distinct objects, we want to choose and **permute** r of them ($r \leq n$). The number of ways to do this is

$$P_{n,r} \equiv \frac{n!}{(n-r)!}.$$

The symbol $P_{n,r}$ is read "the permutation of n things taken r at a time."

Proof. Envision r slots. There are n ways to fill the first slot, $n-1$ ways to fill the second slot, and so on, until we get to the rth slot, in which case there are $n-r+1$ ways to fill it. Thus, by the multiplication rule, there are

$$n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}$$

different permutations. □

Example 1.12. With a group of 5 people, I want to choose a committee with three members: a president, a vice-president, and a secretary. There are

$$P_{5,3} = \frac{5!}{(5-3)!} = \frac{120}{2} = 60$$

different committees possible. **Here, note that order is important.** For any 3 people selected, there are $3! = 6$ different committees possible. □

Example 1.13. In an agricultural experiment, we are examining 10 plots of land; however, only four can be used in an experiment run to test four new (different) fertilizers. How many ways can I choose these four plots and then assign fertilizers?

Solution: There are

$$P_{10,4} = \frac{10!}{(10-4)!} = 5040$$

different permutations. Here, we are assuming fertilizer order is important.

(a) What is the probability of observing the permutation $(7, 4, 2, 6)$?

(b) What is the probability of observing a permutation with only even-numbered plots?

Answers: (a) 1/5040; (b) 120/5040.

CURIOSITY: What happens if the objects to permute are **not distinct**?

Example 1.14. Consider the word *PEPPER*. How many permutations of the letters are possible?

TRICK: Initially, treat all letters as distinct objects by writing, say,

$$P_1E_1P_2P_3E_2R.$$

With $P_1E_1P_2P_3E_2R$, there are $6! = 720$ different orderings of these distinct objects. Now, we recognize that there are

$$\begin{gathered}3! \text{ ways to permute the } P\text{s}\\ 2! \text{ ways to permute the } E\text{s}\\ 1! \text{ ways to permute the } R\text{s}.\end{gathered}$$

Thus, 6! is $3! \times 2! \times 1!$ times **too large**, so we need to divide 6! by $3! \times 2! \times 1!$; i.e., there are

$$\frac{6!}{3!\ 2!\ 1!} = 60$$

possible permutations. □

MULTINOMIAL COEFFICIENTS: Suppose that in a set of n objects, there are n_1 that are similar, n_2 that are similar, ..., n_k that are similar, where $n_1 + n_2 + \cdots + n_k = n$. The number of permutations (i.e., distinguishable permutations, in the sense that the objects are put into distinct groups) of the n objects is given by the **multinomial coefficient**

$$\binom{n}{n_1 n_2 \cdots n_k} \equiv \frac{n!}{n_1!\ n_2!\ \cdots\ n_k!}.$$

NOTE: Multinomial coefficients arise in the algebraic expansion of the multinomial expression $(x_1 + x_2 + \cdots + x_k)$; i.e.,

$$(x_1 + x_2 + \cdots + x_k)^n = \sum_D \frac{n!}{n_1!\ n_2!\ \cdots\ n_k!}\ x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k},$$

where

$$D = \left\{(n_1, n_2, ..., n_k) : \sum_{j=1}^{k} n_i = n\right\}.$$

Example 1.15. How many signals, each consisting of 9 flags in a line, can be made from 4 white flags, 2 blue flags, and 3 yellow flags?

ANSWER:

$$\frac{9!}{4!\;2!\;3!} = 1260.\ \square$$

Example 1.16. In Example 1.15, assuming all permutations are equally-likely, what is the probability that all of the white flags are grouped together? I will offer **two** solutions. The solutions differ in the way I construct the sample space. Define

$$A = \{\text{all four white flags are grouped together}\}.$$

SOLUTION 1. Work with a sample space that does **not** treat the flags as distinct objects, but merely considers color. Then, we know from Example 1.15 that there are 1260 different orderings. Thus,

$$N = \text{number of sample points in } S = 1260.$$

Let n_a denote the number of ways that A can occur. We find n_a by using the multiplication rule.

Stage 1	Pick four adjacent slots	$n_1 = 6$
Stage 2	With the remaining 5 slots, permute the 2 blues and 3 yellows	$n_2 = \frac{5!}{2!3!} = 10$

Thus, $n_a = 6 \times 10 = 60$. Finally, since we have equally likely outcomes, $P(A) = n_a/N = 60/1260 \approx 0.0476$. $\square$

SOLUTION 2. Initially, treat all 9 flags as **distinct objects**; i.e.,

$$W_1W_2W_3W_4B_1B_2Y_1Y_2Y_3,$$

and consider the sample space consisting of the 9! different permutations of these 9 distinct objects. Then,

$$N = \text{number of sample points in } S = 9!$$

Let n_a denote the number of ways that A can occur. We find n_a, again, by using the multiplication rule.

Stage 1	Pick adjacent slots for W_1, W_2, W_3, W_4	$n_1 = 6$
Stage 2	With the four chosen slots, permute W_1, W_2, W_3, W_4	$n_2 = 4!$
Stage 3	With remaining 5 slots, permute B_1, B_2, Y_1, Y_2, Y_3	$n_3 = 5!$

Thus, $n_a = 6 \times 4! \times 5! = 17280$. Finally, since we have equally likely outcomes, $P(A) = n_a/N = 17280/9! \approx 0.0476$. □

1.6.3 Combinations

COMBINATIONS: Given n distinct objects, the number of ways to choose r of them $(r \leq n)$, *without regard to order*, is given by

$$C_{n,r} = \binom{n}{r} \equiv \frac{n!}{r!\,(n-r)!}.$$

The symbol $C_{n,r}$ is read "the combination of n things taken r at a time." By convention, $0! = 1$.

Proof: Choosing r objects is equivalent to breaking the n objects into two distiguishable groups:

Group 1	r chosen
Group 2	$(n-r)$ not chosen.

There are $C_{n,r} = \frac{n!}{r!(n-r)!}$ ways to do this. □

REMARK: We will adopt the notation $\binom{n}{r}$, read "n choose r," as the symbol for $C_{n,r}$. The terms $\binom{n}{r}$ are often called **binomial coefficients** since they arise in the algebraic expansion of a binomial; viz.,

$$(x+y)^n = \sum_{r=0}^{n} \binom{n}{r} x^{n-r} y^r.$$

Example 1.17. Return to Example 1.12. Now, suppose that we only want to choose 3 committee members from 5 (without designations for president, vice-president, and secretary). Then, there are

$$\binom{5}{3} = \frac{5!}{3!\,(5-3)!} = \frac{5 \times 4 \times 3!}{3! \times 2!} = 10$$

different committees. □

NOTE: From Examples 1.12 and 1.17, one should note that

$$P_{n,r} = r! \times C_{n,r}.$$

Recall that combinations do not regard order as important. Thus, once we have chosen our r objects (there are $C_{n,r}$ ways to do this), there are then $r!$ ways to permute those r chosen objects. Thus, we can think of a permutation as simply a combination times the number of ways to permute the r chosen objects.

Example 1.18. A company receives 20 hard drives. Five of the drives will be randomly selected and tested. If all five are satisfactory, the entire lot will be accepted. Otherwise, the entire lot is rejected. If there are really 3 defectives in the lot, what is the probability of accepting the lot?

SOLUTION: First, the number of sample points in S is given by

$$N = \binom{20}{5} = \frac{20!}{5!\,(20-5)!} = 15504.$$

Let A denote the event that the lot is accepted. How many ways can A occur? Use the multiplication rule.

Stage 1	Choose 5 good drives from 17	$\binom{17}{5}$
Stage 2	Choose 0 bad drives from 3	$\binom{3}{0}$

By the multiplication rule, there are $n_a = \binom{17}{5} \times \binom{3}{0} = 6188$ different ways A can occur. Assuming an **equiprobability model** (i.e., each outcome is equally likely), $P(A) = n_a/N = 6188/15504 \approx 0.399$. □

1.7 Conditional probability

MOTIVATION: In some problems, we may be fortunate enough to have **prior knowledge** about the likelihood of events related to the event of interest. It may be of interest to incorporate this information into a probability calculation.

TERMINOLOGY: Let A and B be events in a non-empty sample space S. The **conditional probability** of A, given that B has occurred, is given by

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

provided that $P(B) > 0$.

Example 1.19. A couple has two children.

(a) What is the probability that both are girls?

(b) What is the probability that both are girls, if the eldest is a girl?

SOLUTION: (a) The sample space is given by

$$S = \{(M, M), (M, F), (F, M), (F, F)\}$$

and $N = 4$, the number of sample points in S. Define

$$\begin{aligned} A_1 &= \{\text{1st born child is a girl}\}, \\ A_2 &= \{\text{2nd born child is a girl}\}. \end{aligned}$$

Clearly, $A_1 \cap A_2 = \{(F, F)\}$ and $P(A_1 \cap A_2) = 1/4$, assuming that the four outcomes in S are equally likely. □

SOLUTION: (b) Now, we want $P(A_2|A_1)$. Applying the definition of conditional probability, we get

$$P(A_2|A_1) = \frac{P(A_1 \cap A_2)}{P(A_1)} = \frac{1/4}{1/2} = 1/2. \ \square$$

REMARK: In a profound sense the "new information" in Example 1.19 (i.e., that the eldest is a girl) **induces** a new (or "restricted") sample space given by

$$S^* = \{(F, M), (F, F)\}.$$

On this space, note that $P(A_2) = 1/2$ (computed with respect to S^*). Also note that whether you compute $P(A_2|A_1)$ with the original sample space S or compute $P(A_2)$ with the restricted space S^*, you will get the same answer.

Example 1.20. In a certain community, 36 percent of the families own a dog, 22 percent of the families that own a dog also own a cat, and 30 percent of the families own a cat. A family is selected at random.

(a) Compute the probability that the family owns both a cat and dog.

(b) Compute the probability that the family owns a dog, given that it owns a cat.

SOLUTION: Let $C = \{\text{family owns a cat}\}$ and $D = \{\text{family owns a dog}\}$. In (a), we want $P(C \cap D)$. But,

$$0.22 = P(C|D) = \frac{P(C \cap D)}{P(D)} = \frac{P(C \cap D)}{0.36}.$$

Thus, $P(C \cap D) = 0.36 \times 0.22 = 0.0792$. For (b), simply use the definition of conditional probability:

$$P(D|C) = \frac{P(C \cap D)}{P(C)} = 0.0792/0.30 = 0.264. \ \square$$

PROBABILITY AXIOMS: It is interesting to note that conditional probability satisfies the axioms for a probability set function, when $P(B) > 0$. In particular,

1. $P(A|B) \geq 0$
2. $P(B|B) = 1$
3. If $A_1, A_2, ...$ is a countable sequence of **pairwise mutually exclusive** events (i.e., $A_i \cap A_j = \emptyset$, for $i \neq j$) in S, then

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

MULTIPLICATION LAW OF PROBABILITY: Suppose A and B are events in a non-empty sample space S. Then,

$$\begin{aligned} P(A \cap B) &= P(B|A)P(A) \\ &= P(A|B)P(B). \end{aligned}$$

Proof. As long as $P(A)$ and $P(B)$ are strictly positive, this follows directly from the definition of conditional probability. □

EXTENSION: The multiplication law of probability can be extended to more than 2 events. For example,

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P[(A_1 \cap A_2) \cap A_3] \\ &= P(A_3|A_1 \cap A_2) \times P(A_1 \cap A_2) \\ &= P(A_3|A_1 \cap A_2) \times P(A_2|A_1) \times P(A_1). \end{aligned}$$

NOTE: This suggests that we can compute probabilities like $P(A_1 \cap A_2 \cap A_3)$ "sequentially" by first computing $P(A_1)$, then $P(A_2|A_1)$, then $P(A_3|A_1 \cap A_2)$. The probability of a k-fold intersection can be computed similarly; i.e.,

$$P\left(\bigcap_{i=1}^{k} A_i\right) = P(A_1) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_2) \times \cdots \times P\left(A_k \middle| \bigcap_{i=1}^{k-1} A_i\right).$$

Example 1.21. I am dealt a hand of 5 cards. What is the probability that they are all spades?

Solution. Define A_i to be the event that card i is a spade ($i = 1, 2, 3, 4, 5$). Then,

$$\begin{aligned} P(A_1) &= \frac{13}{52} \\ P(A_2|A_1) &= \frac{12}{51} \\ P(A_3|A_1 \cap A_2) &= \frac{11}{50} \\ P(A_4|A_1 \cap A_2 \cap A_3) &= \frac{10}{49} \\ P(A_5|A_1 \cap A_2 \cap A_3 \cap A_4) &= \frac{9}{48}, \end{aligned}$$

so that

$$P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) = \frac{13}{52} \times \frac{12}{51} \times \frac{11}{50} \times \frac{10}{49} \times \frac{9}{48} \approx 0.0005. \ \square$$

1.8 Independence

TERMINOLOGY: When the occurrence or non-occurrence of A has no effect on whether or not B occurs, and vice-versa, we say that the events A and B are **independent**. Mathematically, we define A and B to be independent iff

$$P(A \cap B) = P(A)P(B).$$

Otherwise, A and B are called **dependent** events. Note that if A and B are independent,

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

and

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

Example 1.22. A red die and a white die are rolled. Let $A = \{4 \text{ on red die}\}$ and $B = \{\text{sum is odd}\}$. Of the 36 outcomes in S, 6 are favorable to A, 18 are favorable to B, and 3 are favorable to $A \cap B$. Thus, since outcomes are assumed to be equally-likely,

$$\frac{3}{36} = P(A \cap B) = P(A)P(B) = \frac{6}{36} \times \frac{18}{36},$$

and the events A and B are independent. □

Example 1.23. In an engineering system, two components are place in a **series**; that is, the system is functional as long as **both** components are. Let A_i; $i = 1, 2$, denote the event that component i is functional. **Assuming independence**, the probability the system is functional is then $P(A_1 \cap A_2) = P(A_1)P(A_2)$. If $P(A_i) = 0.95$, for example, then $P(A_1 \cap A_2) = (0.95)^2 = 0.9025$. □

INDEPENDENCE OF COMPLEMENTS: If A and B are independent events, so are

(a) $\overline{A}$ and B

(b) A and $\overline{B}$

(c) $\overline{A}$ and $\overline{B}$.

Proof. We will only prove (a). The other parts follow similarly.

$P(\overline{A} \cap B) = P(\overline{A}|B)P(B) = [1 - P(A|B)]P(B) = [1 - P(A)]P(B) = P(\overline{A})P(B)$. □

EXTENSION: The concept of independence (and independence of complements) can be extended to any finite number of events in S.

TERMINOLOGY: Let $A_1, A_2, ..., A_n$ denote a collection of $n \geq 2$ events in a non-empty sample space S. The events $A_1, A_2, ..., A_n$ are said to be **mutually independent** if for any subcollection of events, say, $A_{i_1}, A_{i_2}, ..., A_{i_k}$, $2 \leq k \leq n$, we have

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j}).$$

Challenge: Come up with a three events which are **pairwise independent**, but not mutually independent.

COMMON SETTING: Many experiments consist of a sequence of n **trials** that are independent (e.g., flipping a coin 10 times). If A_i denotes the event associated with the ith trial, and the trials are **independent**,

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

Example 1.24. An unbiased die is rolled six times. Let $A_i = \{i \text{ appears on roll } i\}$, for $i = 1, 2, ..., 6$. Then, $P(A_i) = 1/6$, and assuming independence,

$$P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5 \cap A_6) = \prod_{i=1}^{6} P(A_i) = \left(\frac{1}{6}\right)^6.$$

Suppose that if A_i occurs, we will call it “a match.” What is the probability of **at least one** match in the six rolls?

Solution: Let B denote the event that there is **at least one match**. Then, $\overline{B}$ denotes the event that there are no matches. Now,

$$P(\overline{B}) = P(\overline{A}_1 \cap \overline{A}_2 \cap \overline{A}_3 \cap \overline{A}_4 \cap \overline{A}_5 \cap \overline{A}_6) = \prod_{i=1}^{6} P(\overline{A}_i) = \left(\frac{5}{6}\right)^6 = 0.335.$$

Thus, $P(B) = 1 - P(\overline{B}) = 1 - 0.335 = 0.665$, by the complement rule.

Exercise: Generalize this result to an n sided die. What does this probability converge to as $n \to \infty$? □

1.9 Law of Total Probability and Bayes Rule

SETTING: Suppose A and B are events in a non-empty sample space S. We can easily express the event A as follows

$$A = \underbrace{(A \cap B) \cup (A \cap \overline{B})}_{\text{union of disjoint events}}.$$

Thus, by Axiom 3,

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap \overline{B}) \\ &= P(A|B)P(B) + P(A|\overline{B})P(\overline{B}), \end{aligned}$$

where the last step follows from the multiplication law of probability. This is called the **Law of Total Probability** (LOTP). The LOTP can be very helpful. Sometimes computing $P(A|B)$, $P(A|\overline{B})$, and $P(B)$ may be easily computed with available information whereas computing $P(A)$ directly may be difficult.

NOTE: The LOTP follows from the fact that B and $\overline{B}$ **partition** S; that is,

(a) B and $\overline{B}$ are disjoint, and

(b) $B \cup \overline{B} = S$.

Example 1.25. An insurance company classifies people as "accident-prone" and "non-accident-prone." For a fixed year, the probability that an accident-prone person has an accident is 0.4, and the probability that a non-accident-prone person has an accident is 0.2. The population is estimated to be 30 percent accident-prone. (a) What is the probability that a new policy-holder will have an accident?

SOLUTION:

Define $A = \{\text{policy holder has an accident}\}$ and $B = \{\text{policy holder is accident-prone}\}$. Then, $P(B) = 0.3$, $P(A|B) = 0.4$, $P(\overline{B}) = 0.7$, and $P(A|\overline{B}) = 0.2$. By the LOTP,

$$\begin{aligned} P(A) &= P(A|B)P(B) + P(A|\overline{B})P(\overline{B}) \\ &= (0.4)(0.3) + (0.2)(0.7) = 0.26. \ \square \end{aligned}$$

(b) Now suppose that the policy-holder does have an accident. What is the probability that he was "accident-prone?"

SOLUTION: We want $P(B|A)$. Note that

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} = \frac{(0.4)(0.3)}{0.26} = 0.46. \; \Box$$

NOTE: From this last part, we see that, in general,

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

This is a form of **Bayes Rule**.

Example 1.26. A lab test is 95 percent effective in detecting a certain disease when it is present (sensitivity). However, there is a one-percent false-positive rate; that is, the test says that one percent of healthy persons have the disease (specificity). If 0.5 percent of the population truly has the disease, what is the probability that a person has the disease given that

(a) his test is positive?

(b) his test is negative?

SOLUTION: Let $D = \{\text{disease is present}\}$ and $\maltese = \{\text{test is positive}\}$. We are given that $P(D) = 0.005$, $P(\maltese|D) = 0.95$ (sensitivity), $P(\maltese|\overline{D}) = 0.01$ (specificity), and, for (a), we want to compute $P(D|\maltese)$. By Bayes Rule,

$$\begin{aligned} P(D|\maltese) &= \frac{P(\maltese|D)P(D)}{P(\maltese|D)P(D) + P(\maltese|\overline{D})P(\overline{D})} \\ &= \frac{(0.95)(0.005)}{(0.95)(0.005) + (0.01)(0.995)} \approx 0.323. \end{aligned}$$

The reason this is so low is that $P(\maltese|\overline{D})$ is high relative to $P(D)$. In (b), we want $P(D|\overline{\maltese})$. By Bayes Rule,

$$\begin{aligned} P(D|\overline{\maltese}) &= \frac{P(\overline{\maltese}|D)P(D)}{P(\overline{\maltese}|D)P(D) + P(\overline{\maltese}|\overline{D})P(\overline{D})} \\ &= \frac{(0.05)(0.005)}{(0.05)(0.005) + (0.99)(0.995)} \approx 0.00025. \; \Box \end{aligned}$$

Table 1.1: *The general Bayesian scheme.*

Measure before test		Result		Updated measure
$P(D)$		F		$P(D\|F)$
0.005	$\longrightarrow$	$\maltese$	$\longrightarrow$	0.323
0.005	$\longrightarrow$	$\overline{\maltese}$	$\longrightarrow$	0.00025

NOTE: We have discussed the LOTP and Bayes Rule in the case of the partition $\{B, \overline{B}\}$. However, these rules hold for **any** partition.

TERMINOLOGY: A sequence of sets $B_1, B_2, ..., B_k$ is said to form a **partition** of the sample space S if

(a) $B_1 \cup B_2 \cup \cdots \cup B_k = S$ (exhaustive condition), and

(b) $B_i \cap B_j = \emptyset$, for all $i \neq j$ (disjoint condition).

LAW OF TOTAL PROABILITY (*restated*): Suppose that $B_1, B_2, ..., B_k$ forms a partition of S, and suppose $P(B_i) > 0$ for all $i = 1, 2, ..., k$. Then,

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

Proof. Write

$$A = A \cap S = A \cap (B_1 \cup B_2 \cup \cdots \cup B_k) = \bigcup_{i=1}^{k} (A \cap B_i).$$

Thus,

$$P(A) = P\left[\bigcup_{i=1}^{k} (A \cap B_i)\right] = \sum_{i=1}^{k} P(A \cap B_i) = \sum_{i=1}^{k} P(A|B_i)P(B_i). \ \Box$$

BAYES RULE (*restated*): Suppose that $B_1, B_2, ..., B_k$ forms a partition of S, and suppose that $P(A) > 0$ and $P(B_i) > 0$ for all $i = 1, 2, ..., k$. Then,

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

Proof. Simply apply the definition of conditional probability and the multiplication law of probability to get

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

Then, just apply LOTP to $P(A)$ in the denominator to get the result. □

REMARK: Bayesians will call $P(B_j)$ the **prior probability** for the event B_j; they call $P(B_j|A)$ the **posterior probability** of B_j.

Example 1.27. Suppose that a manufacturer buys approximately 60 percent of a raw material (in boxes) from Supplier 1, 30 percent from Supplier 2, and 10 percent from Supplier 3 (these are the prior probabilities). For each supplier, defective rates are as follows: Supplier 1: 0.01, Supplier 2: 0.02, and Supplier 3: 0.03. Suppose that the manufacturer **observes a defective box** of raw material.

(a) What is the probability that it came from Supplier 2?

(b) What is the probability that the defective did **not** come from Supplier 3?

SOLUTION: (a) Let $A = \{\text{observe defective}\}$, and B_1, B_2, and B_3, respectively, denote the events that the box comes from Supplier 1, 2, and 3. Note that $\{B_1, B_2, B_3\}$ partitions the space of possible suppliers. Thus, by Bayes Rule, we have

$$\begin{aligned} P(B_2|A) &= \frac{P(A|B_2)P(B_2)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)} \\ &= \frac{(0.02)(0.3)}{(0.01)(0.6) + (0.02)(0.3) + (0.03)(0.1)} \\ &= 0.40. \end{aligned}$$

SOLUTION: (b) First, compute the posterior probability $P(B_3|A)$. By Bayes Rule,

$$\begin{aligned} P(B_3|A) &= \frac{P(A|B_3)P(B_3)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)} \\ &= \frac{(0.03)(0.1)}{(0.01)(0.6) + (0.02)(0.3) + (0.03)(0.1)} \\ &= 0.20. \end{aligned}$$

Thus, $P(\overline{B}_3|A) = 1 - P(B_3|A) = 1 - 0.20 = 0.80$, by the complement rule. □

2 Discrete Distributions

Complementary reading: Chapter 3 (WMS), except § 3.10-11.

2.1 Random variables

MATHEMATICAL DEFINITION: A **random variable** Y is a function whose domain is the sample space S and whose range is the set of real numbers $\mathcal{R} = \{y : -\infty < y < \infty\}$.

WORKING DEFINITION: A **random variable** is a variable whose observed value is determined by chance.

Example 2.1. Suppose that our experiment consists of flipping two fair coins. The sample space consists of four sample points:

$$S = \{(H,H),(H,T),(T,H),(T,T)\}.$$

Now, let Y denote the number of heads observed. Before we perform the experiment, we do not know, with certainty, the value of Y. What are the **possible values** of Y?

Sample point, E_i	$Y(E_i) = y$
(H,H)	2
(H,T)	1
(T,H)	1
(T,T)	0

In a profound sense, a random variable Y takes sample points $E_i \in S$ and assigns them a **real number**. This is precisely why we can think of Y as a **function**; i.e.,

$$Y[(H,H)] = 2 \qquad Y[(H,T)] = 1 \qquad Y[(T,H)] = 1 \qquad Y[(T,T)] = 0,$$

so that

$$\begin{aligned}
P(Y=2) &= P[\{(H,H)\}] = 1/4 \\
P(Y=1) &= P[\{(H,T)\}] + P[\{(T,H)\}] = 1/4 + 1/4 = 1/2 \\
P(Y=0) &= P[\{(T,T)\}] = 1/4.
\end{aligned}$$

NOTE: From these probability calculations; note that we can

- work on the sample space S and compute probabilities from S, or
- work on $\mathcal{R}$ and compute probabilities for events $\{Y \in B\}$, where $B \subset \mathcal{R}$.

NOTATION: We denote a random variable Y with a **capital letter**; we denote an **observed value** of Y as y, a **lowercase letter**. This is standard notation.

Example 2.2. Let Y denote the weight, in ounces, of the next newborn boy in Columbia, SC. Here, Y is random variable. After the baby is born, we observe $y = 128$. $\square$

2.2 Probability distributions for discrete random variables

TERMINOLOGY: The **support** of a random variable Y is set of all possible values that Y can assume. We will often denote the support set as R. If the random variable Y has a support set R that is either finite or countable, we call Y a **discrete** random variable.

Example 2.3. Suppose that in rolling an unbiased die, we record two random variables:

$$\begin{aligned} X &= \text{face value on the first roll} \\ Y &= \text{number of rolls needed to observe a six.} \end{aligned}$$

The support of X is $R_X = \{1, 2, 3, 4, 5, 6\}$. The support of Y is $R_Y = \{1, 2, 3, ...\}$. R_X is **finite** and R_Y is **countable**; thus, both random variables X and Y are discrete. $\square$

GOAL: With discrete random variables, we would like to assign probabilities to events of the form $\{Y = y\}$. That is, we would like to compute $P(Y = y)$ for any $y \in R$. To do this, one approach is to determine all sample points $E_i \in S$ such that $Y(E_i) = y$ and then compute

$$p_Y(y) \equiv P(Y = y) = \sum P[E_i \in S : Y(E_i) = y],$$

for all $y \in R$. However, as we will see, this approach is often unnecessary.

TERMINOLOGY: The function $p_Y(y) = P(Y = y)$ is called the **probability mass function** (**pmf**) for the discrete random variable Y.

FACTS: The pmf $p_Y(y)$ for a discrete random variable Y consists of **two** parts:

(a) R, the support set of Y

(b) a probability assignment $P(Y = y)$, for all $y \in R$.

PROPERTIES: The pmf $p_Y(y)$ for a discrete random variable Y satisfies the following:

(1) $p_Y(y) > 0$, for all $y \in R$

(2) The **sum** of the probabilities, taken over all support points, must equal one; i.e.,

$$\sum_{y \in R} p_Y(y) = 1.$$

(3) The probability of an event B is computed by adding the probabilities $p_Y(y)$ for all $y \in B$; i.e.,

$$P(Y \in B) = \sum_{y \in B} p_Y(y).$$

Example 2.4. Suppose that we roll an unbiased die twice and observe the face on each roll. Here, the sample space is

$$\begin{aligned} S \;=\; \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\ &(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\ &(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\ &(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}.\end{aligned}$$

Let the random variable Y record the **sum** of the two faces. Here, $R = \{2, 3, ..., 12\}$.

$$\begin{aligned} P(Y = 2) &= P(\{\text{all } E_i \in S \text{ where } Y(E_i) = y = 2\})\\ &= P[\{(1,1)\}] = 1/36. \end{aligned}$$

$$\begin{aligned} P(Y=3) &= P(\{\text{all } E_i \in S \text{ where } Y(E_i) = y = 3\}) \\ &= P[\{(1,2)\}] + P[\{(2,1)\}] = 2/36. \end{aligned}$$

The calculation $P(Y = y)$ is performed similarly for $y = 4, 5, ..., 12$. The pmf for Y can be given as a **formula**, **table**, or **graph**. In tabular form, the pmf of Y is given by

y	2	3	4	5	6	7	8	9	10	11	12
$p_Y(y)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

A **probability histogram** is a display which depicts a pmf in graphical form. The probability histogram for the pmf in Example 2.4 is given in Figure 2.2.

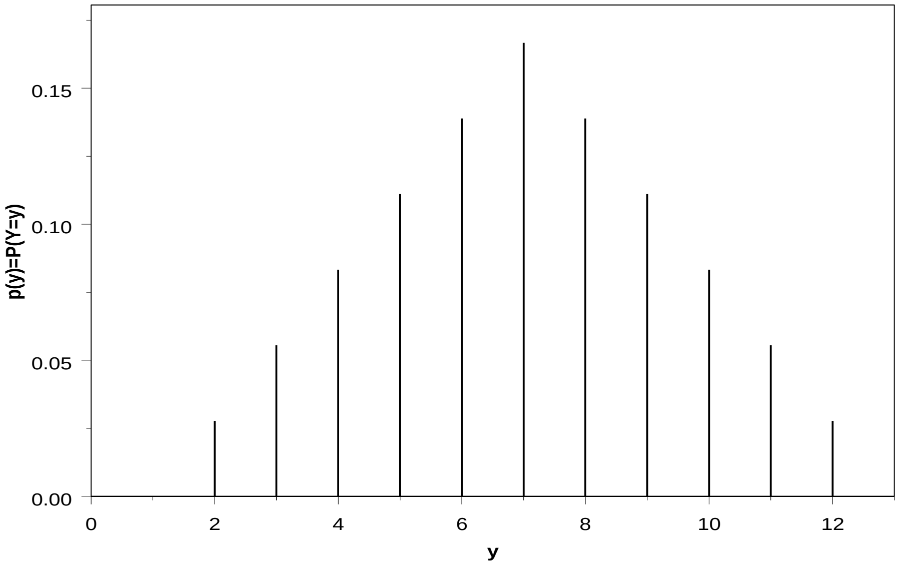

Figure 2.2: *Probability histogram for the pmf in Example* 2.4.

The astute reader will note that a **closed form formula** for the pmf exists; i.e.,

$$p_Y(y) = \begin{cases} \frac{1}{36}\left(6 - |7-y|\right), & y = 2, 3, ..., 12 \\ 0, & \text{otherwise.} \end{cases}$$

Is $p_Y(y)$ **valid**? Yes, since $p_Y(y) > 0$ for all support points $y = 2, 3, ..., 12$, and

$$\sum_{y \in R} p_Y(y) = \sum_{y=2}^{12} \frac{1}{36}\left(6 - |7-y|\right) = 1.$$

QUESTION: Define the events $B_1 = \{\text{the sum is 3}\}$ and $B_2 = \{\text{the sum is odd}\}$. In Example 2.4,

$$P(B_1) = p_Y(3) = 2/36$$

and

$$\begin{aligned} P(B_2) &= \sum_{y \in B_2} p_Y(y) \\ &= p_Y(3) + p_Y(5) + p_Y(7) + p_Y(9) + p_Y(11) \\ &= 2/36 + 4/36 + 6/36 + 4/36 + 2/36 = 1/2. \end{aligned}$$

Example 2.5. An experiment consists of rolling an unbiased die until the first "6" is observed. Let Y denote the number of rolls needed. Here, the support set is $R = \{1, 2, ..., \}$. Assuming independent trials, we have

$$\begin{aligned} P(Y = 1) &= \frac{1}{6} \\ P(Y = 2) &= \frac{5}{6} \times \frac{1}{6} \\ P(Y = 3) &= \frac{5}{6} \times \frac{5}{6} \times \frac{1}{6}; \end{aligned}$$

in general, the probability that y rolls are needed to observe the first "6" is given by

$$P(Y = y) = \frac{1}{6}\left(\frac{5}{6}\right)^{y-1},$$

for all $y = 1, 2, ...$ Thus, the pmf for Y is given by

$$p_Y(y) = \begin{cases} \frac{1}{6}\left(\frac{5}{6}\right)^{y-1}, & y = 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

Is this a **valid** pmf? Clearly, $p_Y(y) > 0$ for all $y \in R$ and

$$\begin{aligned} \sum_{y \in R} p_Y(y) &= \sum_{y=1}^{\infty} \frac{1}{6}\left(\frac{5}{6}\right)^{y-1} \\ &= \sum_{x=0}^{\infty} \frac{1}{6}\left(\frac{5}{6}\right)^{x} \\ &= \left(\frac{\frac{1}{6}}{1 - \frac{5}{6}}\right) = 1. \ \Box \end{aligned}$$

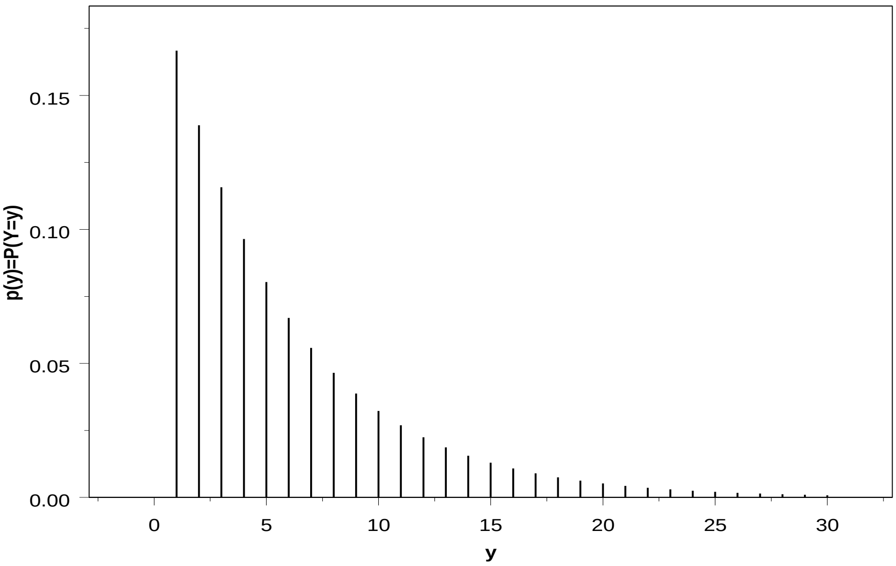

Figure 2.3: *Probability histogram for the pmf in Example* 2.5.

IMPORTANT: In the last calculation, we have used an important fact concerning **infinite geometric series**; namely, if a is any real number and $|r| < 1$. Then,

$$\sum_{x=0}^{\infty} ar^x = \frac{a}{1-r}.$$

The proof of this fact can be found in any standard calculus text. *We will use this fact many times in this course!*

Exercise: In Example 2.5, find $P(B)$, where

$$B = \{\text{the first “6” is observed on an odd-numbered roll}\}.$$

2.3 Mathematical expectation

TERMINOLOGY: Let Y be a discrete random variable with pmf $p_Y(y)$ and support R. The **expected value** of Y is given by

$$E(Y) = \sum_{y \in R} y p_Y(y).$$

DESCRIPTION: In words, the expected value for discrete random variable is a **weighted average** of possible values the variable can assume; each value, y, being weighted with the probability, $p_Y(y)$, that the random variable assumes the corresponding value.

MATHEMATICAL ASIDE: For the expected value $E(Y)$ to exist, the sum above must be **absolutely convergent**; i.e., we need

$$\sum_{y\in R} |y|p_Y(y) < \infty.$$

If $E(Y)$ is not finite; i.e., if $E(Y)=\infty$, we say that $E(Y)$ does not exist.

Example 2.6. Let the random variable Y have pmf

$$p_Y(y) = \begin{cases} \frac{1}{10}(5-y), & y = 1,2,3,4 \\ 0, & \text{otherwise.} \end{cases}$$

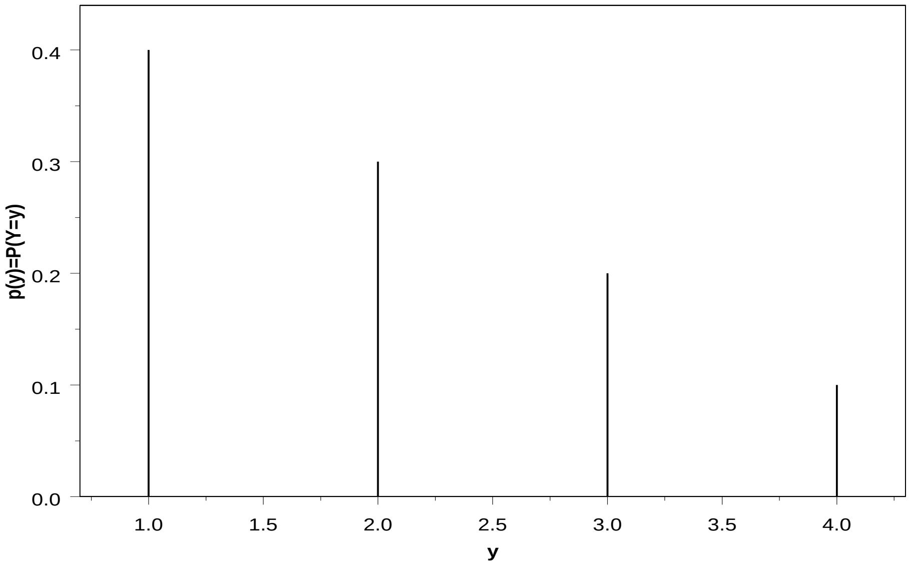

Figure 2.4: *Probability histogram for the pmf in Example* 2.6.

The pmf for Y is depicted in Figure 2.4. The expected value of Y is given by

$$\begin{aligned}\sum_{y\in R} yp_Y(y) &= \sum_{y=1}^{4} y\frac{1}{10}(5-y) \\ &= 1(4/10)+2(3/10)+3(2/10)+4(1/10) = 2. \ \square\end{aligned}$$

Example 2.7. *A random variable whose expected value does not exist.* Suppose that the random variable Y has pmf

$$p_Y(y) = \begin{cases} 1/y, & y \in R \\ 0, & \text{otherwise,} \end{cases}$$

where the support set $R = \{2^i;\ i = 1, 2, 3, ..., \}$. It is easy to see that $p_Y(y)$ is a valid pmf since

$$\sum_{y \in R} p_Y(y) = \sum_{y=1}^{\infty} \left(\frac{1}{2}\right)^y = \left[\sum_{y=0}^{\infty} \left(\frac{1}{2}\right)^y\right] - 1 = \frac{1}{1 - \frac{1}{2}} - 1 = 1.$$

However,

$$E(Y) = \sum_{y \in R} y p_Y(y) = \sum_{y \in R} y \left(\frac{1}{y}\right) = \sum_{y \in R} 1 = \infty,$$

since R, the support set, is countably infinite. □

INTERPRETATION: How is $E(Y)$ interpreted?

(a) the "center of gravity" of a probability distribution

(b) a long-run average

(c) the first **moment** of the random variable.

STATISTICAL CONNECTION: When used in a statistical context, the expected value $E(Y)$ is sometimes called the **mean** of Y, and we might use the symbol μ or μ_Y when discussing it; that is,

$$E(Y) = \mu = \mu_Y.$$

In statistical settings, μ denotes a **population parameter**.

EXPECTATIONS OF FUNCTIONS OF Y: Let Y be a discrete random variable with pmf $p_Y(y)$ and support R, and suppose that g is a real-valued function. Then, $g(Y)$ is a random variable and

$$E[g(Y)] = \sum_{y \in R} g(y) p_Y(y).$$

The proof of this result is given on pp 90 (WMS). □

MATHEMATICAL ASIDE: For the expected value $E[g(Y)]$ to exist, the sum above must be **absolutely convergent**; i.e.,

$$\sum_{y \in R} |g(y)| p_Y(y) < \infty$$

If $E[g(Y)]$ is not finite; i.e., if $E[g(Y)] = \infty$, we say that $E[g(Y)]$ does not exist.

Example 2.8. In Example 2.6, find $E(Y^2)$ and $E(e^Y)$.

Solution: The functions $g_1(Y) = Y^2$ and $g_2(Y) = e^Y$ are real functions of Y. From the definition,

$$\begin{aligned} E(Y^2) &= \sum_{y \in R} y^2 p_Y(y) \\ &= \sum_{y=1}^{4} y^2 \frac{1}{10}(5 - y) \\ &= 1^2(4/10) + 2^2(3/10) + 3^2(2/10) + 4^2(1/10) = 5 \end{aligned}$$

and

$$\begin{aligned} E(e^Y) &= \sum_{y \in R} e^y p_Y(y) \\ &= \sum_{y=1}^{4} e^y \frac{1}{10}(5 - y) \\ &= e^1(4/10) + e^2(3/10) + e^3(2/10) + e^4(1/10) \approx 12.78. \ \Box \end{aligned}$$

Example 2.9. *The discrete uniform distribution.* Suppose that the random variable X has pmf

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

where m is a fixed positive integer larger than 1. Find the expected value of X.

Solution. The expected value of X is given by

$$E(X) = \sum_{x \in R} x p_X(x) = \sum_{x=1}^{m} x \left(\frac{1}{m} \right) = \frac{1}{m} \sum_{x=1}^{m} x = \frac{1}{m} \left[\frac{m(m+1)}{2} \right] = \frac{m+1}{2}.$$

In this calculation, we have used the fact that $\sum_{x=1}^{m} x$, the sum of the first m integers, equals $m(m+1)/2$; this fact can be proven by **mathematical induction**.

REMARK: If $m = 6$, then the discrete uniform distribution serves as a probability model for the outcome of an unbiased die. The expected outcome is $E(X) = \frac{6+1}{2} = 3.5$. □

x	1	2	3	4	5	6
$p_X(x)$	1/6	1/6	1/6	1/6	1/6	1/6

PROPERTIES OF EXPECTATIONS: Let Y be a discrete random variable with pmf $p_Y(y)$ and support R, suppose that $g, g_1, g_2, ..., g_k$ are real-valued functions, and let c be any real constant. Then,

(a) $E(c) = c$

(b) $E[cg(Y)] = cE[g(Y)]$

(c) $E[\sum_{j=1}^k g_j(Y)] = \sum_{j=1}^k E[g_j(Y)]$.

Since $E(\cdot)$ enjoys these above-mentioned properties, we sometimes call E a **linear operator**. Proofs to these facts are easy and are left as exercises.

Example 2.10. In a one-hour period, the number of gallons of a certain toxic chemical that is produced at a local plant, say Y, has the pmf

y	0	1	2	3
$p_Y(y)$	0.2	0.3	0.3	0.2

(a) Compute the expected number of gallons produced during a one-hour period.

(b) The cost (in tens of dollars) to produce Y gallons is given by the cost function $C(Y) = 3 + 12Y + 2Y^2$. What is the expected cost in a one-hour period?

SOLUTION: (a) We have that

$$E(Y) = \sum_{y \in R} y p_Y(y) = 0(0.2) + 1(0.3) + 2(0.3) + 3(0.2) = 1.5.$$

Thus, we would expect 1.5 gallons of the toxic chemical to be produced per hour. For (b), first compute $E(Y^2)$:

$$E(Y^2) = \sum_{y \in R} y^2 p_Y(y) = 0^2(0.2) + 1^2(0.3) + 2^2(0.3) + 3^2(0.2) = 3.3.$$

Now, we use the aforementioned linearity properties to compute

$$\begin{aligned} E[C(Y)] &= E(3 + 12Y + 2Y^2) \\ &= 3 + 12E(Y) + 2E(Y^2) \\ &= 3 + 12(1.5) + 2(3.3) = 27.6. \end{aligned}$$

Thus, the expected hourly cost is \$276.00. □

2.4 Variance

REMARK: We have learned that $E(Y)$ is a measure of the **center** of a probability distribution. Now, we turn our attention to quantifying the **variability** in the distribution.

TERMINOLOGY: Let Y be a discrete random variable with pmf $p_Y(y)$, support R, and mean μ. The **variance** of Y is given by

$$\sigma^2 \equiv V(Y) \equiv E[(Y - \mu)^2] = \sum_{y \in R} (y - \mu)^2 p_Y(y).$$

The **standard deviation** of Y is given by the positive square root of the variance; i.e.,

$$\sigma = \sqrt{V(Y)}.$$

FACTS ABOUT THE VARIANCE:

(a) $\sigma^2 \geq 0$.

(b) $\sigma^2 = 0$ if and only if the random variable Y has a **degenerate distribution**; i.e., all the probability mass is at one point.

(c) The larger (smaller) σ^2 is, the more (less) spread in the possible values of Y about the mean $\mu = E(Y)$.

(d) σ^2 is measured in (units)^2 and σ is measured in the original units.

NOTE: Facts (a), (b), and (c) above are true if we replace σ^2 with σ.

THE VARIANCE COMPUTING FORMULA: Let Y be a random variable (not necessarily a discrete random variable) with pmf $p_Y(y)$ and mean $E(Y) = \mu$. Then

$$V(Y) = E[(Y-\mu)^2] = E(Y^2) - \mu^2.$$

The formula $V(Y) = E(Y^2) - \mu^2$ is called the **variance computing formula**.

Proof. Expand the $(Y-\mu)^2$ term and distribute the expectation operator as follows:

$$\begin{aligned} E[(Y-\mu)^2] &= E(Y^2 - 2\mu Y + \mu^2) \\ &= E(Y^2) - 2\mu E(Y) + \mu^2 \\ &= E(Y^2) - 2\mu^2 + \mu^2 \\ &= E(Y^2) - \mu^2. \ \square \end{aligned}$$

Example 2.11. *The discrete uniform distribution.* Suppose that the random variable X has pmf

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

where m is a fixed positive integer larger than 1. Find the variance of X.

SOLUTION. We will find $\sigma^2 = V(X)$ by using the variance computing formula. In Example 2.9, we computed

$$\mu = E(X) = \frac{m+1}{2}.$$

We first find $E(X^2)$; note that

$$\begin{aligned} E(X^2) = \sum_{x \in R} x^2 p_X(x) = \sum_{x=1}^{m} x^2 \left(\frac{1}{m}\right) = \frac{1}{m}\sum_{x=1}^{m} x^2 &= \frac{1}{m}\left[\frac{m(m+1)(2m+1)}{6}\right] \\ &= \frac{(m+1)(2m+1)}{6}. \end{aligned}$$

Above, we have used the fact that $\sum_{x=1}^{m} x^2$, the sum of the first m squared integers, equals $m(m+1)(2m+1)/6$; this fact can be proven by **mathematical induction**. The variance of X is equal to

$$\begin{aligned}\sigma^2 &= E(X^2) - \mu^2 \\ &= \frac{(m+1)(2m+1)}{6} - \left(\frac{m+1}{2}\right)^2 \\ &= \frac{m^2-1}{12}.\end{aligned}$$

Note that if $m = 6$, as for our unbiased die example, $\sigma^2 = 35/12$. □

EXERCISE: Find σ^2 for the pmf in Example 2.6 (notes).

IMPORTANT RESULT: Let Y be a random variable (not necessarily a discrete random variable) and suppose that a and b are real constants. Then

$$V(a + bY) = b^2V(Y).$$

Proof. Exercise. □

REMARK: Taking $b = 0$ above, we see that $V(a) = 0$, for any constant a. This makes sense intuitively. The variance is a measure of variability for a random variable; a constant (such as a) does not vary. Also, by taking $a = 0$, we see that $V(bY) = b^2V(Y)$. Both of these facts are important and we will use them repeatedly.

2.5 Moment generating functions

TERMINOLOGY: Let Y be a discrete random variable with pmf $p_Y(y)$ and support R. The **moment generating function (mgf)** for Y, denoted by $m_Y(t)$, is given by

$$m_Y(t) = E(e^{tY}) = \sum_{y \in R} e^{ty}p_Y(y),$$

provided $E(e^{tY}) < \infty$ for t in an open neighborhood about 0; i.e., there exists some $h > 0$ such that $E(e^{tY}) < \infty$ for all $t \in (-h, h)$. If $E(e^{tY})$ does not exist in an open neighborhood of 0, we say that the moment generating function does not exist.

TERMINOLOGY: We call $E(Y^k)$ the k**th moment** of the random variable Y.

$$\begin{array}{ll} E(Y) & \text{1st moment (mean!)} \\ E(Y^2) & \text{2nd moment} \\ E(Y^3) & \text{3rd moment} \\ \vdots & \quad\vdots \end{array}$$

NOTATION: WMS use the notation μ'_k to denote the kth moment; i.e., $E(Y^k) = \mu'_k$. This is common notation in statistics applications, but I rarely use it.

REMARK: The moment generating function (mgf) can be used to generate moments. In fact, from the theory of Laplace transforms, it follows that if the mgf exists, it characterizes an infinite set of moments. So, how do we **generate** moments?

RESULT: Let Y denote a random variable (not necessarily a discrete random variable) with support R and mgf $m_Y(t)$. Then,

$$E(Y^k) = \left.\frac{d^k m_Y(t)}{dt^k}\right|_{t=0}.$$

Note that derivatives are taken with respect to t.

Proof. Assume, without loss, that Y is discrete. With $k = 1$, we have

$$\begin{aligned} \frac{d}{dt}m_Y(t) &= \frac{d}{dt}\sum_{y\in R} e^{ty}p_Y(y) \\ &= \sum_{y\in R}\frac{d}{dt}e^{ty}p_Y(y) = \sum_{y\in R} ye^{ty}p_Y(y) = E(Ye^{tY}). \end{aligned}$$

Thus, it follows that

$$\left.\frac{dm_Y(t)}{dt}\right|_{t=0} = E(Ye^{tY})\Big|_{t=0} = E(Y).$$

Continuing to take higher-order derivatives, we can prove that

$$\left.\frac{d^k m_Y(t)}{dt^k}\right|_{t=0} = E(Y^k),$$

for any integer $k \geq 1$. Thus, the result follows. □

MATHEMATICAL ASIDE: In the second line of the proof of the last result, we interchanged the derivative and (possibly infinite) sum. This is permitted as long as $m_Y(t) = E(e^{tY})$ exists.

COMPUTING MEANS AND VARIANCES: Let Y denote a random variable (not necessarily a discrete random variable) with mgf $m_Y(t)$. Then, we know that

$$E(Y) = \left.\frac{dm_Y(t)}{dt}\right|_{t=0},$$

and

$$E(Y^2) = \left.\frac{d^2m_Y(t)}{dt^2}\right|_{t=0}.$$

Thus,

$$\begin{aligned} V(Y) &= E(Y^2) - [E(Y)]^2 \\ &= \left.\frac{d^2m_Y(t)}{dt^2}\right|_{t=0} - \left[\left.\frac{dm_Y(t)}{dt}\right|_{t=0}\right]^2 \\ &\equiv m''_Y(0) - [m'_Y(0)]^2. \end{aligned}$$

REMARK: In many applications, being able to compute means and variances is important. **Thus, we can use the mgf as a tool to do this.** This is helpful because sometimes computing

$$E(Y) = \sum_{y \in R} y p_Y(y)$$

directly (or even higher order moments) may be extremely difficult, depending on the form of $p_Y(y)$.

Example 2.12. Suppose that Y is a random variable with pmf

$$p_Y(y) = \begin{cases} \left(\frac{1}{2}\right)^y, & y = 1, 2, 3, ... \\ 0, & \text{otherwise.} \end{cases}$$

Find the mean of Y.

SOLUTION. Using the definition of expected values, the mean of Y is given by

$$E(Y) = \sum_{y \in R} y p_Y(y) = \sum_{y=1}^{\infty} y \left(\frac{1}{2}\right)^y.$$

Finding this infinite sum is quite difficult (at least, this sum is not a geometric sum). It is easier to use moment generating functions! The mgf of Y is given by

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \sum_{y \in R} e^{ty} p_Y(y) \\ &= \sum_{y=1}^{\infty} e^{ty} \left(\frac{1}{2}\right)^y \\ &= \sum_{y=1}^{\infty} \left(\frac{e^t}{2}\right)^y \\ &= \left[\sum_{y=0}^{\infty} \left(\frac{e^t}{2}\right)^y\right] - 1 = \frac{1}{1-\frac{e^t}{2}} - 1 = \frac{e^t}{2-e^t}, \end{aligned}$$

for values of $t < \ln 2$ (why?). Thus,

$$\begin{aligned} E(Y) = \left.\frac{dm_Y(t)}{dt}\right|_{t=0} &= \left.\frac{d}{dt}\left(\frac{e^t}{2-e^t}\right)\right|_{t=0} \\ &= \left.\frac{e^t(2-e^t) - e^t(-e^t)}{(2-e^t)^2}\right|_{t=0} = 2. \ \Box \end{aligned}$$

Example 2.13. Let the random variable Y have pmf $p_Y(y)$ given by

$$p_Y(y) = \begin{cases} \frac{1}{6}(3-y), & y = 0, 1, 2 \\ 0, & \text{otherwise.} \end{cases}$$

For this probability distribution, simple calculations (verify!) show that

$$\begin{aligned} E(Y) &= 2/3 \\ V(Y) &= 5/9. \end{aligned}$$

Let's "check" these calculations using the mgf. It is given by

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \sum_{y \in R} e^{ty} p_Y(y) \\ &= e^{t(0)}\frac{3}{6} + e^{t(1)}\frac{2}{6} + e^{t(2)}\frac{1}{6} \\ &= \frac{3}{6} + \frac{2}{6}e^t + \frac{1}{6}e^{2t}. \end{aligned}$$

Taking derivatives of $m_Y(t)$ with respect to t, we get

$$\frac{d}{dt}m_Y(t) = \frac{2}{6}e^t + \frac{2}{6}e^{2t}$$

and

$$\frac{d^2}{dt^2}m_Y(t) = \frac{2}{6}e^t + \frac{4}{6}e^{2t}.$$

Thus,

$$E(Y) = \left.\frac{dm_Y(t)}{dt}\right|_{t=0} = \frac{2}{6}e^0 + \frac{2}{6}e^{2(0)} = 4/6 = 2/3$$

$$E(Y^2) = \left.\frac{d^2m_Y(t)}{dt^2}\right|_{t=0} = \frac{2}{6}e^0 + \frac{4}{6}e^{2(0)} = 1$$

so that

$$V(Y) = E(Y^2) - [E(Y)]^2 = 1 - (2/3)^2 = 5/9.$$

So, in this example, we can use the mgf to get $E(Y)$ and $V(Y)$, or we can compute $E(Y)$ and $V(Y)$ directly. We get the same answer, as we should. □

REMARK: Not only is the mgf a tool for computing moments, but it also helps us to **characterize** a probability distribution. How? When an mgf exists, it happens to be **unique**. Thus, if two random variables have same mgf, then they have the **same** probability distribution! Sometimes, this is referred to as the **uniqueness property** of mgfs (it is based on the uniqueness of Laplace transforms). For now, however, it suffices to envision the mgf as a "special expectation" that generates moments. This, in turn, helps us to compute means and variances of random variables.

2.6 Binomial distribution

BERNOULLI TRIALS: Many experiments consist of a sequence of trials, where

(i) each trial results in a "success" or a "failure,"

(ii) the trials are **independent**, and

(iii) the probability of "success," denoted by p, $0 < p < 1$, is the **same** on every trial.

TERMINOLOGY: In a sequence of n Bernoulli trials, denote by Y the **number of successes** (out of n, where n is fixed). We call Y a **binomial** random variable, and say that "Y has a **binomial distribution** with parameters n and success probability p." Shorthand notation is $Y \sim b(n, p)$.

Example 2.14. Each of the following situations represent **binomial experiments**. (Are you satisfied with the Bernoulli assumptions in each instance?)

(a) Suppose we flip a fair coin 10 times and let Y denote the number of tails in 10 flips. Here, $Y \sim b(n = 10, p = 0.5)$.

(b) In an agricultural experiment, forty percent of all plots respond to a certain treatment. I have four plots of land to be treated. If Y is the number of plots that respond to the treatment, then $Y \sim b(n = 4, p = 0.4)$.

(c) In rural Kenya, the prevalence rate for HIV is estimated to be around 8 percent. Let Y denote the number of HIV infecteds in a sample of 740 individuals. Here, $Y \sim b(n = 740, p = 0.08)$.

(d) It is known that screws produced by a certain company do not meet specifications (i.e., are defective) with probability 0.001. Let Y denote the number of defectives in a package of 40. Then, $Y \sim b(n = 40, p = 0.001)$. □

DERIVATION: We now derive the pmf of a binomial random variable. That is, we need to compute $p_Y(y) = P(Y = y)$, for each possible value of $y \in R$. Recall that Y is the number of "successes" in n Bernoulli trials so the support set is $R = \{y : y = 0, 1, 2, ..., n\}$.

QUESTION: In a sequence of n trials, how can we get **exactly** y successes? Denoting

$$S = \text{success}$$
$$F = \text{failure}$$

a possible sample point may be

$$SSFSFSFFS \cdots FSF.$$

Because the trials are **independent**, the probability that we get any particular ordering of y successes and $n-y$ failures is $p^y(1-p)^{n-y}$. Now, how many ways are there to choose y successes from n trials? We know that there are $\binom{n}{y}$ ways to do this. Thus, the pmf for Y is, for $0 < p < 1$,

$$p_Y(y) = \begin{cases} \binom{n}{y}p^y(1-p)^{n-y}, & y = 0, 1, 2, ..., n \\ 0, & \text{otherwise.} \end{cases}$$

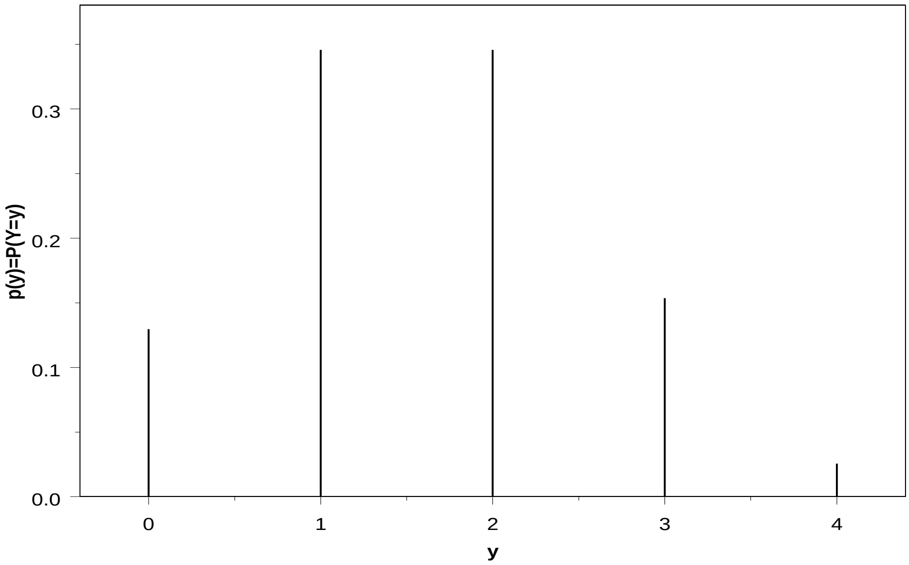

Figure 2.5: *Probability histogram for the number of plots which respond to treatment. This represents the* $b(n = 4, p = 0.4)$ *model in Example* 2.14(b).

Example 2.15. In Example 2.14(b), assume that $Y \sim b(n = 4, p = 0.4)$. Here are the probability calculations for this binomial model:

$$\begin{aligned} P(Y = 0) = p_Y(0) &= \tbinom{4}{0}(0.4)^0(1-0.4)^{4-0} = 1 \times (0.4)^0 \times (0.6)^4 = 0.1296 \\ P(Y = 1) = p_Y(1) &= \tbinom{4}{1}(0.4)^1(1-0.4)^{4-1} = 4 \times (0.4)^1 \times (0.6)^3 = 0.3456 \\ P(Y = 2) = p_Y(2) &= \tbinom{4}{2}(0.4)^2(1-0.4)^{4-2} = 6 \times (0.4)^2 \times (0.6)^2 = 0.3456 \\ P(Y = 3) = p_Y(3) &= \tbinom{4}{3}(0.4)^3(1-0.4)^{4-3} = 4 \times (0.4)^3 \times (0.6)^1 = 0.1536 \\ P(Y = 4) = p_Y(4) &= \tbinom{4}{4}(0.4)^4(1-0.4)^{4-4} = 1 \times (0.4)^4 \times (0.6)^0 = 0.0256 \end{aligned}$$

The probability histogram is depicted in Figure 2.5. □

Example 2.16. In a small clinical trial with 20 patients, let Y denote the number of patients that respond to a new skin rash treatment. The physicians assume that a binomial model is appropriate so that $Y \sim b(n = 20, p)$, where p denotes the probability of response to the treatment. In a statistical setting, p would be an unknown **parameter** that we desire to **estimate**. For this problem, we'll assume that $p = 0.4$. Compute (a) $P(Y = 5)$, (b) $P(Y \geq 5)$, and (c) $P(Y < 10)$.

(a) $P(Y = 5) = p_Y(5) = \binom{20}{5}(0.4)^5(0.6)^{20-5} = 0.0746$.

(b)

$$P(Y \geq 5) = \sum_{y=5}^{20} P(Y = y) = \sum_{y=5}^{20} \binom{20}{y}(0.4)^y(0.6)^{20-y}$$

This computation involves using the binomial pmf 16 times and adding the results! TRICK: Instead of computing the sum $\sum_{y=5}^{20}\binom{20}{y}(0.4)^y(0.6)^{20-y}$ directly, we can write

$$P(Y \geq 5) = 1 - P(Y \leq 4),$$

by the complement rule. We do this because WMS's Appendix III (Table 1, pp. 783-785) contains binomial probability calculations of the form

$$F_Y(a) \equiv P(Y \leq a) = \sum_{y=0}^{a} \binom{n}{y} p^y(1-p)^{n-y},$$

for different n and p. With $n = 20$ and $p = 0.4$, we see from Table 1 that

$$P(Y \leq 4) = 0.051.$$

Thus, $P(Y \geq 5) = 1 - 0.051 = 0.949$.

(c) $P(Y < 10) = P(Y \leq 9) = 0.755$, from Table 1. $\square$

REMARK: The function

$$F_Y(y) \equiv P(Y \leq y)$$

is called the **cumulative distribution function**; we'll talk more about this function in the next chapter.

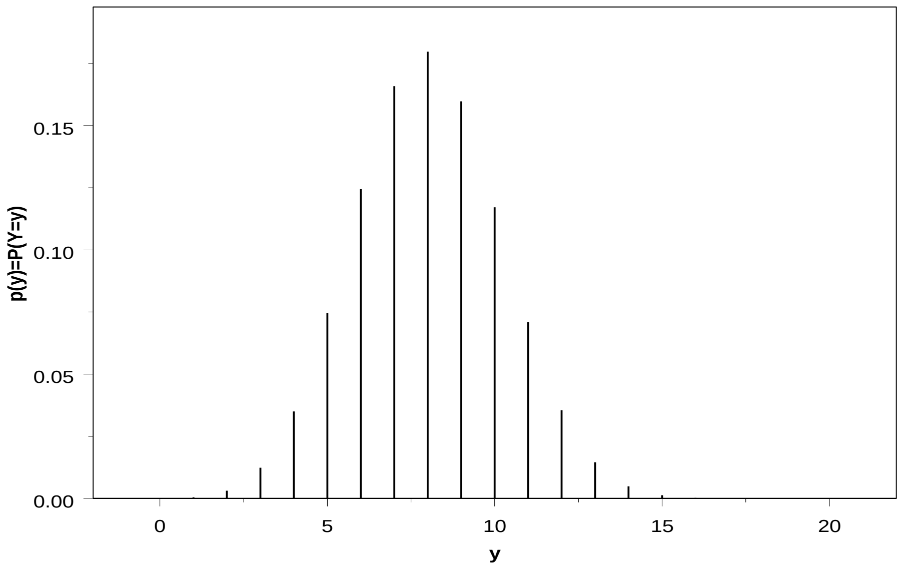

Figure 2.6: *Probability histogram for the number of patients responding to treatment. This represents the* $b(n = 20, p = 0.4)$ *model in Example* 2.16.

CURIOSITY: Is the binomial pmf a **valid** pmf? Clearly $p_Y(y) > 0$ for all y. To check that the pmf sums to one, consider the **binomial expansion**

$$[p + (1 - p)]^n = \sum_{y=0}^{n} \binom{n}{y} p^y (1 - p)^{n-y}.$$

The LHS clearly equals 1, and the RHS represents the $b(n, p)$ pmf. Thus, $p_Y(y)$ is valid.

MGF FOR THE BINOMIAL DISTRIBUTION: Suppose that $Y \sim b(n, p)$. Then the mgf of Y is given by

$$m_Y(t) = E(e^{tY}) = \sum_{y=0}^{n} e^{ty} \binom{n}{y} p^y (1 - p)^{n-y} = \sum_{y=0}^{n} \binom{n}{y} (pe^t)^y (1 - p)^{n-y} = (q + pe^t)^n,$$

where $q = 1 - p$. The last step follows from noting that $\sum_{y=0}^{n} \binom{n}{y} (pe^t)^y (1 - p)^{n-y}$ is the **binomial expansion** of $(q + pe^t)^n$. □

MEAN AND VARIANCE OF THE BINOMIAL DISTRIBUTION: We want to compute $E(Y)$ and $V(Y)$ where $Y \sim b(n, p)$. To do this, we will use the mgf. Taking the derivative of $m_Y(t)$ with respect t, we get

$$m'_Y(t) \equiv \frac{d}{dt}m_Y(t) = \frac{d}{dt}(q+pe^t)^n = n(q+pe^t)^{n-1}pe^t.$$

Thus,

$$E(Y) = \frac{d}{dt}m_Y(t)\bigg|_{t=0} = n(q+pe^0)^{n-1}pe^0 = n(q+p)^{n-1}p = np,$$

since $q+p=1$. Now, we need to find the second moment. By using the product rule for derivatives, we have

$$\frac{d^2}{dt^2}m_Y(t) = \frac{d}{dt}\underbrace{n(q+pe^t)^{n-1}pe^t}_{m'_Y(t)} = n(n-1)(q+pe^t)^{n-2}(pe^t)^2 + n(q+pe^t)^{n-1}pe^t.$$

Thus,

$$E(Y^2) = \frac{d^2}{dt^2}m_Y(t)\bigg|_{t=0} = n(n-1)(q+pe^0)^{n-2}(pe^0)^2 + n(q+pe^0)^{n-1}pe^0 = n(n-1)p^2+np.$$

Finally, the variance is calculated by appealing to the variance computing formula; i.e.,

$$\begin{aligned} V(Y) &= E(Y^2) - [E(Y)]^2 \\ &= n(n-1)p^2 + np - (np)^2 \\ &= np(1-p). \ \square \end{aligned}$$

Example 2.17. Artichokes are a marine climate vegetable and thrive in the cooler coastal climates. Most will grow on a wide range of soils, but produce best on a deep, fertile, well-drained soil. Suppose that 15 artichoke seeds are planted in identical soils and temperatures, and let Y denote the number of seeds that germinate. If 60 percent of all seeds germinate (on average) and we assume a $b(15, 0.6)$ probability model for Y, the mean number of seeds that will germinate is

$$E(Y) = np = 15(0.6) = 9.$$

The variance is

$$\sigma^2 = np(1-p) = 15(0.6)(0.4) = 3.6 \text{ (seeds)}^2$$

The standard deviation is

$$\sigma = \sqrt{3.6} \approx 1.9 \text{ seeds. } \square$$

SPECIAL BINOMIAL DISTRIBUTION: In the $b(n, p)$ family, when $n = 1$, the binomial pmf reduces to

$$p_Y(y) = \begin{cases} p^y(1-p)^{1-y}, & y = 0, 1 \\ 0, & \text{otherwise.} \end{cases}$$

This is sometimes called the **Bernoulli distribution**. Shorthand notation is $Y \sim b(1, p)$. The sum of n independent $b(1, p)$ random variables actually follows a $b(n, p)$ distribution!

2.7 Geometric distribution

TERMINOLOGY: Imagine an experiment where Bernoulli trials are observed. If Y denotes the trial on which the **first success** occurs, then Y is said to follow a **geometric distribution** with parameter p, the probability of success on any one trial, $0 < p < 1$. This is sometimes written as $Y \sim \text{geom}(p)$. The pmf for Y is given by

$$p_Y(y) = \begin{cases} (1-p)^{y-1}p, & y = 1, 2, 3, ... \\ 0, & \text{otherwise.} \end{cases}$$

RATIONALE: The form of this pmf makes intuitive sense; we need $y - 1$ failures (each of which occurs with probability $1 - p$), and then a success on the yth trial (this occurs with probability p). By independence, we multiply

$$\underbrace{(1-p) \times (1-p) \times \cdots \times (1-p)}_{y-1 \text{ failures}} \times p = (1-p)^{y-1}p.$$

NOTE: Clearly $p_Y(y) > 0$ for all y. Does $p_Y(y)$ sum to one? Note that

$$\begin{aligned} \sum_{y=1}^{\infty}(1-p)^{y-1}p &= p\sum_{x=0}^{\infty}(1-p)^x \\ &= \frac{p}{1-(1-p)} = 1. \end{aligned}$$

In the last step, we realized that $\sum_{x=0}^{\infty}(1-p)^x$ is an **infinite geometric sum** with common ratio $1 - p$. □

Example 2.18. Biology students are checking the eye color of fruit flies. For each fly, the probability of observing white eyes is $p = 0.25$. What is the probability the **first** white-eyed fly will be observed among the first five flies that we check?

SOLUTION: Let Y denote the number of flies needed to observe the first white-eyed fly. We need to compute $P(Y \leq 5)$. We can envision each fly as a Bernoulli trial (each fly either has white eyes or not). If we assume that the flies are independent, then a geometric model is appropriate; i.e., $Y \sim \text{geom}(p = 0.25)$, so that

$$\begin{aligned}
P(Y = 1) = p_Y(1) = (1 - 0.25)^{1-1}(0.25) &= 0.25 \\
P(Y = 2) = p_Y(2) = (1 - 0.25)^{2-1}(0.25) &\approx 0.19 \\
P(Y = 3) = p_Y(3) = (1 - 0.25)^{3-1}(0.25) &\approx 0.14 \\
P(Y = 4) = p_Y(4) = (1 - 0.25)^{4-1}(0.25) &\approx 0.11 \\
P(Y = 5) = p_Y(5) = (1 - 0.25)^{5-1}(0.25) &\approx 0.08.
\end{aligned}$$

Thus, $P(Y \leq 5) = \sum_{y=1}^{5} P(Y = y) \approx 0.77$. The pmf for the geom$(p = 0.25)$ model is depicted in Figure 2.7. □

MGF FOR THE GEOMETRIC DISTRIBUTION: Suppose that $Y \sim \text{geom}(p)$. Then the mgf of Y is given by

$$m_Y(t) = \frac{pe^t}{1 - qe^t},$$

where $q = 1 - p$, for $t < -\ln q$.

Proof. Exercise. □

MEAN AND VARIANCE OF THE GEOMETRIC DISTRIBUTION: With the mgf, we can derive the mean and variance. Differentiating the mgf, we get

$$m'_Y(t) \equiv \frac{d}{dt} m_Y(t) = \frac{d}{dt}\left(\frac{pe^t}{1 - qe^t}\right) = \frac{pe^t(1 - qe^t) - pe^t(-qe^t)}{(1 - qe^t)^2}.$$

Thus,

$$E(Y) = \frac{d}{dt} m_Y(t)\bigg|_{t=0} = \frac{pe^0(1 - qe^0) - pe^0(-qe^0)}{(1 - qe^0)^2} = \frac{p(1 - q) - p(-q)}{(1 - q)^2} = \frac{1}{p}.$$

Similar calculations show

$$E(Y^2) = \frac{d^2}{dt^2} m_Y(t)\bigg|_{t=0} = \frac{1 + q}{p^2}.$$

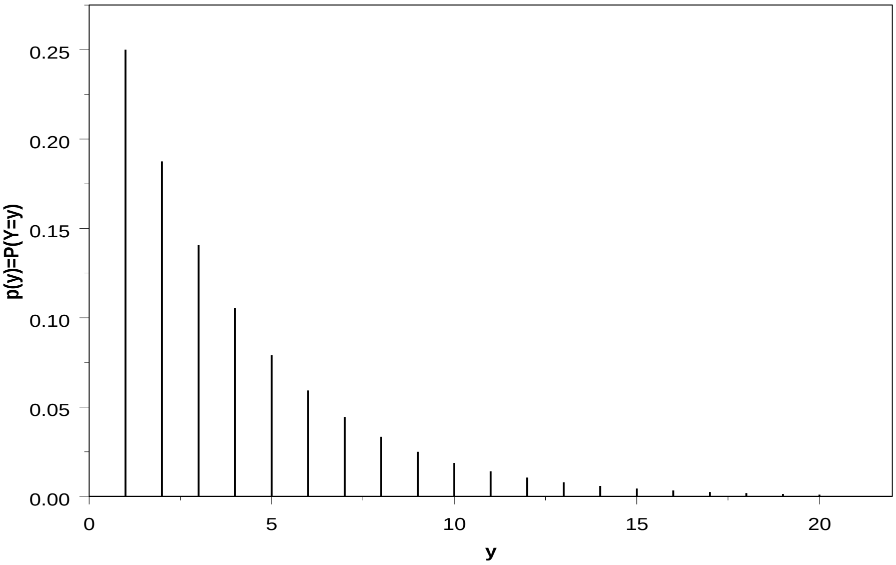

Figure 2.7: *Probability histogram for the number of flies needed to find the first white-eyed fly. This represents the* geom($p = 0.25$) *model in Example* 2.18.

Finally,

$$\begin{aligned} V(Y) &= E(Y^2) - [E(Y)]^2 \\ &= \frac{1+q}{p^2} - \left(\frac{1}{p}\right)^2 \\ &= \frac{q}{p^2}. \ \square \end{aligned}$$

Example 2.19. At an apple orchard in Maine, bags of "20-lbs" are continually observed until the first underweight bag is discovered. Suppose that four percent of bags are under filled. If we assume that the bags are independent, and if Y denotes the the number of bags observed, then $Y \sim \text{geom}(p = 0.04)$. The mean number of bags we will observe is

$$E(Y) = \frac{1}{p} = \frac{1}{0.04} = 25 \text{ bags}.$$

The variance is

$$V(Y) = \frac{q}{p^2} = \frac{0.96}{(0.04)^2} = 600 \ (\text{bags})^2. \ \square$$

2.8 Negative binomial distribution

NOTE: The negative binomial distribution can be motivated from two perspectives:

- as a generalization of the geometric
- as a "reversal" of the binomial.

Recall that the geometric random variable was defined to be the number of trials needed to observe the **first success** in a sequence of Bernoulli trials.

TERMINOLOGY: Imagine an experiment where Bernoulli trials are observed. If Y denotes the trial on which the r**th success** occurs, $r \geq 1$, then Y has a **negative binomial distribution** with parameters r and p, where p denotes the probability of success on any one trial, $0 < p < 1$. This is sometimes written as $Y \sim \text{nib}(r, p)$.

PMF FOR THE NEGATIVE BINOMIAL: The pmf for $Y \sim \text{nib}(r, p)$ is given by

$$p_Y(y) = \begin{cases} \binom{y-1}{r-1} p^r (1-p)^{y-r}, & y = r, r+1, r+2, ... \\ 0, & \text{otherwise.} \end{cases}$$

Of course, when $r = 1$, the $\text{nib}(r, p)$ pmf reduces to the $\text{geom}(p)$ pmf.

RATIONALE: The logic behind the form of $p_Y(y)$ is as follows. If the rth success occurs on the yth trial, then $r-1$ successes must have occurred during the 1st $y-1$ trials. The total number of sample points (in the underlying sample space S) where this is the case is given by the binomial coefficient $\binom{y-1}{r-1}$, which counts the number of ways you order $r-1$ successes and $y-r$ failures in the 1st $y-1$ trials. The probability of any particular ordering, by independence, is given by $p^{r-1}(1-p)^{y-r}$. Now, on the yth trial, we observe the rth success (this occurs with probability p). Thus, putting it all together, we get

$$\underbrace{\binom{y-1}{r-1} p^{r-1}(1-p)^{y-r}}_{\text{pertains to 1st } y-1 \text{ trials}} \times p = \binom{y-1}{r-1} p^r (1-p)^{y-r}.$$

Example 2.20. A botanist in Iowa City is observing oak trees for the presence of a certain disease. From past experience, it is known that 30 percent of all trees are infected ($p = 0.30$). Treating each tree as a Bernoulli trial (i.e., each tree is infected/not), what is the probability that she will observe the 3rd infected tree ($r = 3$) on the 6th or 7th observed tree?

SOLUTION. Let Y denote the tree on which she observes the 3rd infected tree. Then, $Y \sim \text{nib}(r = 3, p = 0.3)$. We want to compute $P(Y = 6 \text{ or } Y = 7)$.

$$P(Y=6) = \binom{6-1}{3-1}(0.3)^3(1-0.3)^{6-3} = 0.0926$$

$$P(Y=7) = \binom{7-1}{3-1}(0.3)^3(1-0.3)^{7-3} = 0.0972$$

Thus,

$$\begin{aligned} P(Y = 6 \text{ or } Y = 7) &= P(Y=6) + P(Y=7) \\ &= 0.0926 + 0.0972 \\ &= 0.1898. \ \square \end{aligned}$$

RELATIONSHIP WITH THE BINOMIAL: Recall that in a binomial experiment, we **fix the number of Bernoulli trials**, n, and we observe the number of successes. However, in a negative binomial experiment, we **fix the number of successes** we are to observe, r, and we continue to observe Bernoulli trials until we reach that success. This is another way to think about the negative binomial model.

MGF FOR THE NEGATIVE BINOMIAL DISTRIBUTION: Suppose that $Y \sim \text{nib}(r, p)$. The mgf of Y is given by

$$m_Y(t) = \left(\frac{pe^t}{1 - qe^t}\right)^r,$$

where $q = 1 - p$, for all $t < -\ln q$. Before we prove this, let's state and prove a lemma.

LEMMA. Suppose that r is a nonnegative integer. Then,

$$\sum_{y=r}^{\infty} \binom{y-1}{r-1}(qe^t)^{y-r} = (1 - qe^t)^{-r}.$$

Proof of lemma. Consider the function $f(w) = (1-w)^{-r}$, where r is a nonnegative integer. It is easy to show that

$$\begin{aligned} f'(w) &= r(1-w)^{-(r+1)} \\ f''(w) &= r(r+1)(1-w)^{-(r+2)} \\ &\vdots \end{aligned}$$

In general, $f^{(z)}(w) = r(r+1)\cdots(r+z-1)(1-w)^{-(r+z)}$, where $f^{(z)}(w)$ denotes the zth derivative of f with respect to w. Note that

$$f^{(z)}(w)\Big|_{w=0} = r(r+1)\cdots(r+z-1).$$

Now, consider writing the McLaurin Series expansion of $f(w)$; i.e., a Taylor Series expansion of $f(w)$ about $w=0$; this expansion is given by

$$\begin{aligned} f(w) &= \sum_{z=0}^{\infty} \frac{f^{(z)}(0)w^z}{z!} \\ &= \sum_{z=0}^{\infty} \frac{r(r+1)\cdots(r+z-1)}{z!} w^z \\ &= \sum_{z=0}^{\infty} \binom{z+r-1}{r-1} w^z. \end{aligned}$$

Now, letting $w = qe^t$ and $z = y-r$, the lemma is proven for $0<q<1$. □

MGF: Now that we are finished with the lemma, let's find the mgf of the nib(r,p) random variable. With $q = 1-p$,

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \sum_{y=r}^{\infty} e^{ty}\binom{y-1}{r-1} p^r q^{y-r} \\ &= \sum_{y=r}^{\infty} e^{t(y-r)}e^{tr}\binom{y-1}{r-1} p^r q^{y-r} \\ &= (pe^t)^r \sum_{y=r}^{\infty} \binom{y-1}{r-1} (qe^t)^{y-r} \\ &= (pe^t)^r(1-qe^t)^{-r} \\ &= \left(\frac{pe^t}{1-qe^t}\right)^r, \end{aligned}$$

for $t < -\ln q$, where the penultimate step follows from the lemma. □

REMARK: Showing that the nib(r, p) distribution sums to one can be done by using a similar series expansion as above. We omit it for brevity.

MEAN AND VARIANCE OF THE NEGATIVE BINOMIAL DISTRIBUTION: For a nib(r, p) random variable, with $q = 1 - p$,

$$E(Y) = \frac{r}{p}$$

and

$$V(Y) = \frac{rq}{p^2}.$$

Proof. Exercise. □

2.9 Hypergeometric distribution

SETTING: Consider a collection of N objects (e.g., people, poker chips, plots of land, etc.) and suppose that we have **two** dichotomous classes, Class 1 and Class 2. For example, the objects and classes might be

Poker chips	red/blue
People	infected/not infected
Plots of land	respond to treatment/not

From the collection of N objects, we observe a sample of $n < N$ of them, and record Y, the number of objects in Class 1 (i.e., the number of "successes").

REMARK: This sounds like binomial setup! However, the difference is that, here, N, the **population size**, is finite (the population size, theoretically, is assumed to be infinite in the binomial model). Thus, if we sample from a population of objects **without replacement**, then the "success" probability changes trial to trial! This, violates the binomial model assumptions!! Of course, if N is large (i.e., in very large populations), the two models will be similar, because the change in the probability of success from trial to trial will be small (maybe so small that it is not of practical concern).

HYPERGEOMETRIC DISTRIBUTION: Envision a collection of n objects sampled (at random and without replacement) from a population of size N, where r denotes the size of Class 1 and $N-r$ denotes the size of Class 2. Let Y denote the number of objects in the sample that belong to Class 1. Then, Y has a **hypergeometric distribution**, written $Y \sim \text{hyper}(N, n, r)$, where

$$\begin{aligned} N &= \text{total number of objects} \\ r &= \text{number of the 1st class (e.g., ``success'')} \\ N-r &= \text{number of the 2nd class (e.g., ``failure'')} \\ n &= \text{number of objects sampled.} \end{aligned}$$

HYPERGEOMETRIC PMF: The pmf for $Y \sim \text{hyper}(N, n, r)$ is given by

$$p_Y(y) = \begin{cases} \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}, & y \in R \\ 0, & \text{otherwise,} \end{cases}$$

where the support set $R = \{y \in \mathcal{N} : \max(0, n-N+r) \leq y \leq \min(n, r)\}$.

BREAKDOWN: In the hyper(N, n, r) pmf, we have three parts:

$$\begin{aligned} \binom{r}{y} &= \text{number of ways to choose } y \text{ Class 1 objects from } r \\ \binom{N-r}{n-y} &= \text{number of ways to choose } n-y \text{ Class 2 objects from } N-r \\ \binom{N}{n} &= \text{number of sample points.} \end{aligned}$$

REMARK: In the hypergeometric model, it follows that $p_Y(y)$ sums to 1 over the support R, but we omit this proof for brevity (see Exercise 3.176, pp 148, WMS).

Example 2.21. In my fish tank at home, there are 50 fish. Ten have been tagged. If I catch 7 fish (and random, and without replacement), what is the probability that **exactly** two are tagged?

SOLUTION. Here, $N = 50$ (total number of fish), $n = 7$ (sample size), $r = 10$ (tagged fish; Class 1), $N - r = 40$ (untagged fish; Class 2), and $y = 2$ (number of tagged fish caught). Thus,

$$P(Y = y) = P(Y = 2) = p_Y(2) = \frac{\binom{10}{2}\binom{40}{5}}{\binom{50}{7}} = 0.2964.$$

What about the probability that my catch contains **at most two** tagged fish?

SOLUTION. Here, we want

$$\begin{aligned} P(Y \leq 2) &= P(Y=0)+P(Y=1)+P(Y=2) \\ &= \frac{\binom{10}{0}\binom{40}{7}}{\binom{50}{7}} + \frac{\binom{10}{1}\binom{40}{6}}{\binom{50}{7}} + \frac{\binom{10}{2}\binom{40}{5}}{\binom{50}{7}} \\ &= 0.1867 + 0.3843 + 0.2964 = 0.8674. \ \Box \end{aligned}$$

Example 2.22. A supplier ships parts to another company in lots of 25 parts. The receiving company has an **acceptance sampling plan** which adopts the following acceptance rule:

> "....sample 5 parts at random and without replacement. If there are no defectives in the sample, accept the entire lot; otherwise, reject the entire lot."

Let Y denote the number of defectives in the sampled parts (i.e., out of 5). Then, $Y \sim \text{hyper}(25, 5, r)$, where r denotes the number defectives in the lot (in real life, r is unknown). Define

$$OC(p) = P(Y=0) = \frac{\binom{r}{0}\binom{25-r}{5}}{\binom{25}{5}},$$

where $p = r/25$ denotes the true proportion of defectives in the lot. The symbol $OC(p)$ denotes the **probability of accepting the lot** (which is a function of p). Consider the following table, whose entries are computed using the above probability expression:

r	p	$OC(p)$
0	0	1.00
1	0.04	0.80
2	0.08	0.63
3	0.12	0.50
4	0.16	0.38
5	0.20	0.29
10	0.40	0.06
15	0.60	0.01

REMARK: The graph of $OC(p)$ versus p is sometimes called an **operating characteristic curve**. Of course, as r (or equivalently, p) increases, the probability of accepting the lot decreases. Acceptance sampling is important in **statistical process control** used in engineering and manufacturing settings. In practice, lot sizes may be very large (e.g., $N = 1000$, etc.), and developing sound sampling plans is crucial in order to avoid using defective parts in finished products. □

MEAN AND VARIANCE OF THE HYPERGEOMETRIC DISTRIBUTION: If $Y \sim$ hyper(N, n, r), then

$$E(Y) = n\left(\frac{r}{N}\right)$$

and

$$V(Y) = n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right).$$

We will prove this result later in the course.

RELATIONSHIP WITH THE BINOMIAL: As noted earlier, the binomial and hypergeometric models are similar. The key difference is that in a binomial experiment, p does not change from trial to trial, but it does in the hypergeometric setting, noticeably if N is small. However, one can show that, for y fixed,

$$\lim_{N\to\infty} \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}} = \underbrace{\binom{n}{y} p^y (1-p)^{n-y}}_{b(n,p)\text{ pmf}}$$

as $r/N \to p$. The upshot is this: if N is large (i.e., the population size is large), a binomial probability calculation, with $p = r/N$, closely **approximates** the corresponding hypergeometric probability calculation. See pp 123 (WMS).

Example 2.23. In a small town, there are 900 right-handed individuals and 100 left-handed individuals. We take a sample of size $n = 20$ individuals from this town (at random and without replacement). What is the probability that 4 or more people in the sample are left-handed?

SOLUTION. Let X denote the number of left-handed individuals in our sample. Let's compute this probability $P(X \geq 4)$ using both the binomial and hypergeometric models.

- **Hypergeometric**: Here, $N = 1000$, $r = 100$, $N - r = 900$, and $n = 20$. Thus,

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - \sum_{x=0}^{3} \frac{\binom{100}{x}\binom{900}{20-x}}{\binom{1000}{20}} \approx 0.130947.$$

- **Binomial**: Here, $n = 20$ and $p = r/N = 0.10$. Thus,

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - \sum_{x=0}^{3} \binom{20}{x}(0.1)^x(0.9)^{20-x} \approx 0.132953. \ \square$$

REMARK: Of course, since the binomial and hypergeometric models are similar when N is large, their means and variances are similar too. Note the similarities; recall that the quantity $r/N \to p$, as $N \to \infty$.

$$E(Y) = n\left(\frac{r}{N}\right) \approx np$$

and

$$V(Y) = n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right) \approx np(1-p).$$

2.10 Poisson distribution

TERMINOLOGY: Let the number of occurrences in a given continuous interval of time or space be counted. A **Poisson process** enjoys the following properties:

(1) the number of occurrences in non-overlapping intervals are **independent** random variables.

(2) The probability of an occurrence in a sufficiently short interval is **proportional to the length** of the interval.

(3) The probability of 2 or more occurrences in a sufficiently short interval is zero.

GOAL: Suppose that an experiment satisfies the above three conditions, and let Y denote the number of occurrences in an interval of length one. Our goal is to find an expression for $p_Y(y) = P(Y = y)$, the pmf of Y.

APPROACH: Envision partitioning the unit interval $[0, 1]$ into n subintervals, each of size $\frac{1}{n}$. Now, if n is sufficiently large (i.e., much larger than y), then we can approximate the probability that y events occur in this unit interval by finding the probability that exactly one event (occurrence) occurs in exactly y of the subintervals.

- By Property (2), we know that the probability of one event in any one subinterval is **proportional** to the subinterval's length, say λ/n, where λ is the proportionality constant.
- By Property (3), the probability of more than one occurrence in any subinterval is **zero** (for n large).
- Consider the occurrence/non-occurrence of an event in each subinterval as a **Bernoulli trial**. Then, by Property (1), we have a sequence of n Bernoulli trials, each with probability of "success" $p = \lambda/n$. Thus, a binomial calculation gives
$$P(Y = y) \approx \binom{n}{y}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y}.$$

Now, to get a better approximation, we let n grow without bound. Then,

$$\begin{aligned}\lim_{n\to\infty} P(Y = y) &= \lim_{n\to\infty} \frac{n!}{y!\,(n-y)!}\lambda^y\left(\frac{1}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^n\left(\frac{1}{1-\frac{\lambda}{n}}\right)^y \\ &= \lim_{n\to\infty} \underbrace{\frac{n(n-1)\cdots(n-y+1)}{n^y}}_{a_n}\underbrace{\frac{\lambda^y}{y!}}_{b_n}\underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{c_n}\underbrace{\left(\frac{1}{1-\frac{\lambda}{n}}\right)^y}_{d_n}\end{aligned}$$

Now, the limit of the product is the product of the limits. Thus,

$$\begin{aligned}\lim_{n\to\infty} a_n &= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-y+1)}{n^y} = 1 \\ \lim_{n\to\infty} b_n &= \lim_{n\to\infty} \frac{\lambda^y}{y!} = \frac{\lambda^y}{y!} \\ \lim_{n\to\infty} c_n &= \lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda} \\ \lim_{n\to\infty} d_n &= \lim_{n\to\infty} \left(\frac{1}{1-\frac{\lambda}{n}}\right)^y = 1.\end{aligned}$$

Thus,

$$p_Y(y) = \begin{cases} \frac{\lambda^y e^{-\lambda}}{y!}, & y = 0, 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

This is the pmf of a **Poisson** random variable with parameter λ. We sometimes write $Y \sim \text{Poisson}(\lambda)$. That $p_Y(y)$ sums to one is easily seen as

$$\begin{aligned} \sum_{y \in R} p_Y(y) &= \sum_{y=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!} \\ &= e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \\ &= e^{-\lambda} e^{\lambda} = 1, \end{aligned}$$

since $e^{\lambda} = \sum_{y=0}^{\infty} \lambda^y / y!$, the McLaurin series expansion of e^{λ}. □

EXAMPLES OF POISSON PROCESSES:

(1) counting the number of people in a certain community living to 100 years of age.

(2) counting the number of customers entering a post office in a given day.

(3) counting the number of α-particles discharged from a radioactive substance in a fixed period of time.

(4) counting the number of blemishes on a piece of artificial turf.

(5) counting the number of chocolate chips in a Chips-Ahoy cookie.

Example 2.24. The number of cars abandoned weekly on a certain highway is modeled using a Poisson distribution with $\lambda = 2.2$. In a given week, what is the probability that

(a) no cars are abandoned?

(b) exactly one car is abandoned?

(c) at most one car is abandoned?

(d) at least one car is abandoned?

SOLUTIONS. Let Y denote the number of cars abandoned weekly.

(a)

$$P(Y=0)=p_Y(0)=\frac{(2.2)^0e^{-2.2}}{0!}=e^{-2.2}=0.1108$$

(b)

$$P(Y=1)=p_Y(1)=\frac{(2.2)^1e^{-2.2}}{1!}=2.2e^{-2.2}=0.2438$$

(c) $P(Y\leq 1)=P(Y=0)+P(Y=1)=p_Y(0)+p_Y(1)=0.1108+0.2438=0.3456$

(d) $P(Y\geq 1)=1-P(Y=0)=1-p_Y(0)=1-0.1108=0.8892$. □

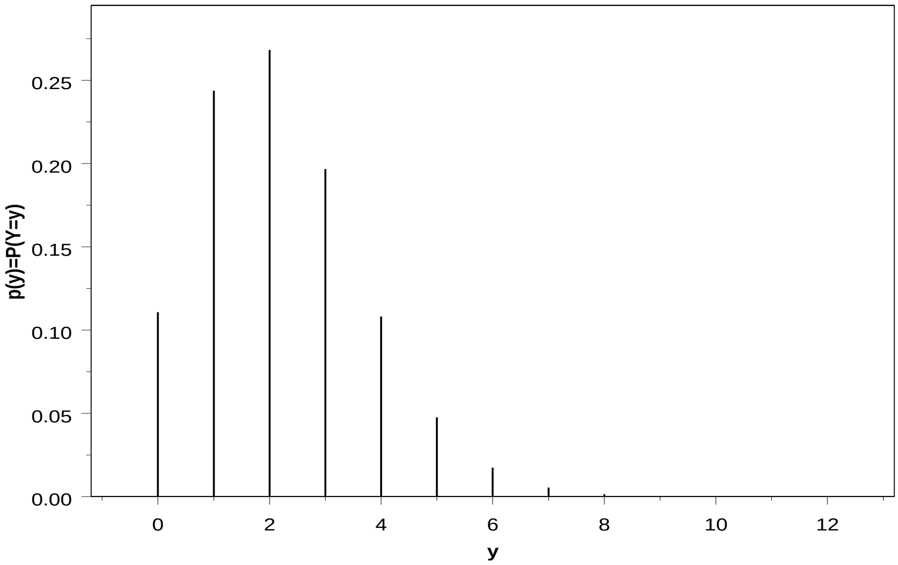

Figure 2.8: *Probability histogram for the number of abandoned cars. This represents the* Poisson($\lambda=2.2$) *model in Example* 2.24.

REMARK: WMS's Appendix III, (Table 3, pp 787-791) includes an impressive table for Poisson probabilities of the form

$$F_Y(a)=P(Y\leq a)=\sum_{y=0}^{a}\frac{\lambda^y e^{-\lambda}}{y!}.$$

Recall that this function is called the **cumulative distribution function** of Y. This makes computing compound event probabilities much easier.

MGF FOR THE POISSON DISTRIBUTION: Suppose that $Y \sim \text{Poisson}(\lambda)$. The mgf of Y, for all t, is given by

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \sum_{y=0}^{\infty} e^{ty} \frac{\lambda^y e^{-\lambda}}{y!} \\ &= e^{-\lambda} \underbrace{\sum_{y=0}^{\infty} \frac{(\lambda e^t)^y}{y!}}_{e^{\lambda e^t}} \\ &= e^{-\lambda} e^{\lambda e^t} = \exp[\lambda(e^t - 1)]. \end{aligned}$$

MEAN AND VARIANCE OF THE POISSON DISTRIBUTION: With the mgf, we can derive the mean and variance. Differentiating the mgf, we get

$$m'_Y(t) \equiv \frac{d}{dt} m_Y(t) = \frac{d}{dt} \exp[\lambda(e^t - 1)] = \lambda e^t \exp[\lambda(e^t - 1)].$$

Thus,

$$E(Y) = \frac{d}{dt} m_Y(t)\bigg|_{t=0} = \lambda e^0 \exp[\lambda(e^0 - 1)] = \lambda.$$

Now, we need to find the second moment. By using the product rule for derivatives, we have

$$\frac{d^2}{dt^2} m_Y(t) = \frac{d}{dt} \underbrace{\lambda e^t \exp[\lambda(e^t - 1)]}_{m'_Y(t)} = \lambda e^t \exp[\lambda(e^t - 1)] + (\lambda e^t)^2 \exp[\lambda(e^t - 1)].$$

Thus, $E(Y^2) = \lambda + \lambda^2$ and

$$\begin{aligned} V(Y) &= E(Y^2) - [E(Y)]^2 \\ &= \lambda + \lambda^2 - \lambda^2 \\ &= \lambda. \end{aligned}$$

REVELATION: With a Poisson model, the mean and variance are **always** equal. □

Example 2.25. Suppose that Y denotes the number of monthly defects observed at an automotive plant. From past experience, engineers believe the Poisson model is appropriate and that $Y \sim \text{Poisson}(7)$.

QUESTION 1: What is the probability that, in any given month, we observe 11 or more defectives?

Solution. We want to compute

$$P(Y \geq 11) = 1 - \underbrace{P(Y \leq 10)}_{\text{Table 3}} = 1 - 0.901 = 0.099.$$

Question 2: What about the probability that, in a given year, we have two or more months with 11 or more defectives?

Solution. First, we assume that the 12 months are independent (is this reasonable?), and call the event $B = \{11 \text{ or more defects in a month}\}$ a "success." Thus, under our independence assumptions and viewing each month as a "trial," we have a sequence of 12 Bernoulli trials with "success" probability $p = P(B) = 0.099$. Let X denote the number of months where we observe 11 or more defects. Then, $X \sim b(12, 0.099)$, and

$$\begin{aligned}
P(X \geq 2) &= 1 - P(X=0) - P(X=1) \\
&= 1 - \binom{12}{0}(0.099)^0(1-0.099)^{12} - \binom{12}{1}(0.099)^1(1-0.099)^{11} \\
&= 1 - 0.2862 - 0.3774 = 0.3364. \ \square
\end{aligned}$$

POISSON PROCESSES OF ARBITRARY LENGTH: If events or occurrences in a Poisson process occur at a rate of λ per unit time or space, then the number of occurrences in an interval of length t also follows a Poisson distribution with mean λt.

Example 2.26. Phone calls arrive at a switchboard according to a Poisson process, at a rate of $\lambda = 3$ per minute. Thus, if Y represents the number of calls received in 5 minutes, we have that $Y \sim \text{Poisson}(15)$. The probability that 8 or fewer calls come in during a 5-minute span is given by

$$P(Y \leq 8) = \sum_{y=0}^{8} \frac{15^y e^{-15}}{y!} = 0.037,$$

from Table 3. $\square$

POISSON-BINOMIAL LINK: We have seen that the hypergeometric and binomial models are related; as it turns out, so are the Poisson and binomial models. This should not be surprising because we derived the Poisson pmf by appealing to a binomial approximation.

RELATIONSHIP: Suppose that $Y \sim b(n,p)$. If n is large and p is small, then

$$p_Y(y) = \binom{n}{y} p^y (1-p)^{n-y} \approx \frac{\lambda^y e^{-\lambda}}{y!},$$

for $y \in R = \{0, 1, 2, ..., n\}$, where $\lambda = np$.

Example 2.27. Hepatitis C (HCV) is a viral infection that causes cirrhosis and cancer of the liver. Since HCV is transmitted through contact with infectious blood, screening donors is important to prevent further transmission. The World Health Organization has projected that HCV will be a major burden on the US health care system before the year 2020. For public-health reasons, researchers take a sample of $n = 1875$ blood donors and screen each individual for HCV. If 3 percent of the entire population is infected, what is the probability that 50 or more are HCV-positive?

Solution. Let Y denote the number of HCV-infected individuals in our sample. We compute this probability $P(Y \geq 50)$ using both the binomial and Poisson models.

- **Binomial**: Here, $n = 1875$ and $p = 0.03$. Thus,

$$P(Y \geq 50) = \sum_{y=50}^{1875} \binom{1875}{y} (0.03)^y (0.97)^{1875-y} \approx 0.818783.$$

- **Poisson**: Here, $\lambda = np = 1875(0.03) \approx 56.25$. Thus,

$$P(Y \geq 50) = \sum_{y=50}^{\infty} \frac{(56.25)^y e^{-56.25}}{y!} \approx 0.814932.$$

As we can see, the Poisson approximation is quite good. □

RELATIONSHIP: One can see that the hypergeometric, binomial, and Poisson models are related in the following way:

$$\text{hyper}(N, n, r) \longleftrightarrow b(n,p) \longleftrightarrow \text{Poisson}(\lambda)$$

The first link results when N is large and $r/N \rightarrow p$. The second link results when n is large and p is small so that $\lambda/n \rightarrow p$. When these situations are combined, as you might suspect, one can approximate the hypergeometric model with a Poisson model!!

3 Continuous Distributions

Complementary reading from WMS: Chapter 4 (omit § 4.11).

3.1 Introduction

RECALL: In the last chapter, we focused on **discrete** random variables. Recall that a discrete random variable is one that can assume only a finite or countable number of values. We also learned about **probability mass functions** (**pmfs**). Loosely speaking, these were functions that told us how to assign probabilities and to which points we assign probabilities.

TERMINOLOGY: A random variable is said to be **continuous** if its support set is uncountable (i.e., the random variable can assume an uncountably infinite number of values). We will present an alternate definition shortly.

3.2 Cumulative distribution functions

NEW: We now introduce a new function associated with any random variable (discrete or continuous).

TERMINOLOGY: The **cumulative distribution function (cdf)** of a random variable Y, denoted by $F_Y(y)$, is given by

$$F_Y(y) = P(Y \leq y), \text{ for all } y \in \mathcal{R}.$$

Note that the cdf is defined for all $y \in \mathcal{R}$, not just for those values of $y \in R$, the support set of Y.

REMARK: Every random variable, discrete or continuous, has a cdf. We'll start by computing some cdfs for **discrete** random variables.

Example 3.1. Let the random variable Y have pmf

$$p_Y(y) = \begin{cases} \frac{1}{6}(3-y), & y = 0,1,2 \\ 0, & \text{otherwise.} \end{cases}$$

Consider the following probability calculations:

$$\begin{aligned} F_Y(0) &= P(Y \leq 0) = P(Y = 0) = \frac{3}{6} \\ F_Y(1) &= P(Y \leq 1) = P(Y = 0) + P(Y = 1) = \frac{3}{6} + \frac{2}{6} = \frac{5}{6} \\ F_Y(2) &= P(Y \leq 2) = P(Y = 0) + P(Y = 1) + P(Y = 2) = \frac{3}{6} + \frac{2}{6} + \frac{1}{6} = 1. \end{aligned}$$

Furthermore,

- for any $y < 0$, $P(Y \leq y) = 0$
- for any $0 < y < 1$, $P(Y \leq y) = P(Y = 0) = \frac{3}{6}$
- for any $1 < y < 2$, $P(Y \leq y) = P(Y = 0) + P(Y = 1) = \frac{3}{6} + \frac{2}{6} = \frac{5}{6}$
- for any $y > 2$, $P(Y \leq y) = P(Y = 0) + P(Y = 1) + P(Y = 2) = \frac{3}{6} + \frac{2}{6} + \frac{1}{6} = 1.$

Putting this all together, we get

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ \frac{3}{6}, & 0 \leq y < 1 \\ \frac{5}{6}, & 1 \leq y < 2 \\ 1, & y \geq 2. \end{cases}$$

Note that we have defined $F_Y(y)$ for all $y \in \mathcal{R}$. Some points are worth mentioning concerning the graphs of the **pmf** and **cdf**:

- **PMF**
  - The height of the bar above y is the probability that Y assumes that value.
  - For any y not equal to $0, 1$, or 2, $p_Y(y) = 0$.

- **CDF**
  - $F_Y(y)$ is a nondecreasing function; see theoretical properties below.
  - $0 \leq F_Y(y) \leq 1$; this makes sense since $F_Y(y)$ is a probability!!
  - The height of the "jump" at a particular point is equal to the probability associated with that point.

THEORETICAL PROPERTIES: Let Y be a random variable (discrete or continuous) and suppose that $F_Y(y)$ is the cdf for Y. Then

(i) $\lim_{y\to-\infty} F_Y(y) = 0$,

(ii) $\lim_{y\to+\infty} F_Y(y) = 1$,

(iii) $F_Y(y)$ is a **right continuous** function; that is, for any real a, $\lim_{y\to a^+} F_Y(y) = F_Y(a)$, and

(iv) $F_Y(y)$ is a **non-decreasing** function; that is, for any $y_1 \leq y_2$, $F_Y(y_1) \leq F_Y(y_2)$.

EXERCISE: Graph the cdf for the $b(5, 0.2)$ and Poisson(2) distributions.

3.3 Continuous random variables

ALTERNATE DEFINITION: A random variable is said to be **continuous** if its cdf $F_Y(y)$ is a continuous function of y.

RECALL: The cdfs associated with discrete random variables are **step-functions**. Such functions are certainly not continuous; however, they are still right continuous.

TERMINOLOGY: Let Y be a continuous random variable with cdf $F_Y(y)$. The **probability density function** (**pdf**) for Y, denoted by $f_Y(y)$, is given by

$$f_Y(y) = \frac{d}{dy}F_Y(y),$$

provided that $\frac{d}{dy}F_Y(y) \equiv F'_Y(y)$ exists. Furthermore, appealing to the Fundamental Theorem of Calculus, we know that

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt.$$

REMARK: These equations illustrate key relationships linking pdfs and cdfs for **continuous** random variables!!

PROPERTIES OF CONTINUOUS PDFs: Suppose that Y is a continuous random variable with pdf $f_Y(y)$ and support R. Then

(1) $f_Y(y) > 0$, for all $y \in R$;

(2) $\int_R f_Y(y)dy = 1$; i.e., the total area under the pdf equals one;

(3) The probability of an event B is computed by integrating the pdf $f_Y(y)$ over B; i.e., $P(Y \in B) = \int_B f_Y(y)dy$, for any $B \subset \mathcal{R}$.

REMARK: Compare these to the analogous results for the **discrete** case (see page 28 in the notes). The only difference is that in the continuous case, integrals replace sums.

Example 3.2. Suppose that Y has the pdf

$$f_Y(y) = \begin{cases} \frac{1}{2}, & 0 < y < 2 \\ 0, & \text{otherwise.} \end{cases}$$

This pdf is depicted in Figure 3.9. We want to find the cdf $F_Y(y)$. To do this, we need to compute $F_Y(y) = P(Y \leq y)$ for all $y \in \mathcal{R}$. There are three cases:

- when $y \leq 0$, we have
$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt = \int_{-\infty}^{y} 0dt = 0;$$
- when $0 < y < 2$, we have
$$\begin{aligned} F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt &= \int_{-\infty}^{0} 0dt + \int_{0}^{y} \frac{1}{2}dt \\ &= 0 + \frac{t}{2}\Big|_0^y = y/2; \end{aligned}$$

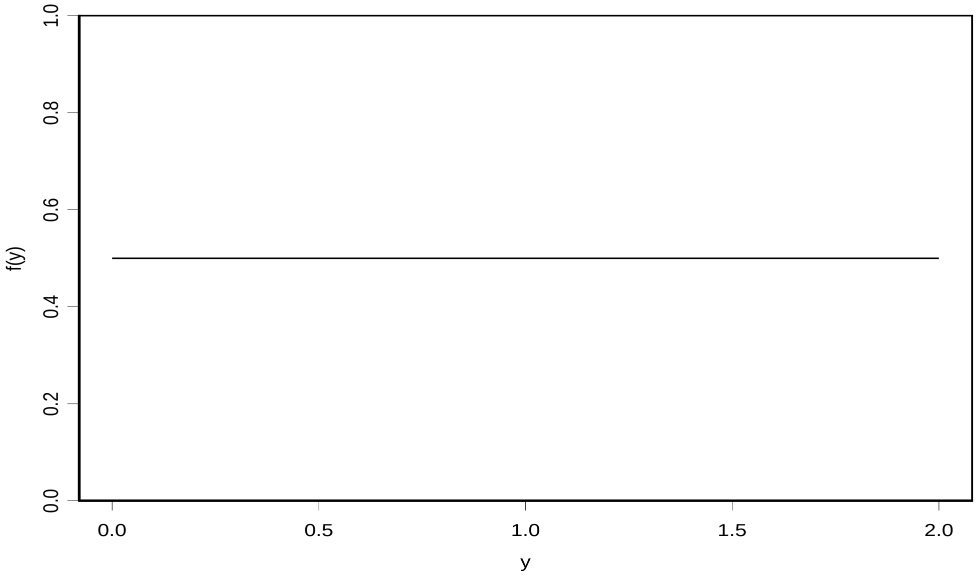

Figure 3.9: *Probability density function, $f_Y(y)$, in Example* 3.2.

- when $y \geq 2$, we have

$$\begin{aligned} F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt &= \int_{-\infty}^{0} 0dt + \int_{0}^{2} \frac{1}{2}dt + \int_{2}^{y} 0dt \\ &= 0 + \left.\frac{t}{2}\right|_{0}^{2} + 0 = 1. \end{aligned}$$

Putting it all together, we have

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ y/2, & 0 \leq y < 2 \\ 1, & y \geq 2. \end{cases}$$

Example 3.3. Remission times for a certain group of leukemia patients (Y, measured in months) has cdf

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ 1 - e^{-y/3}, & y \geq 0. \end{cases}$$

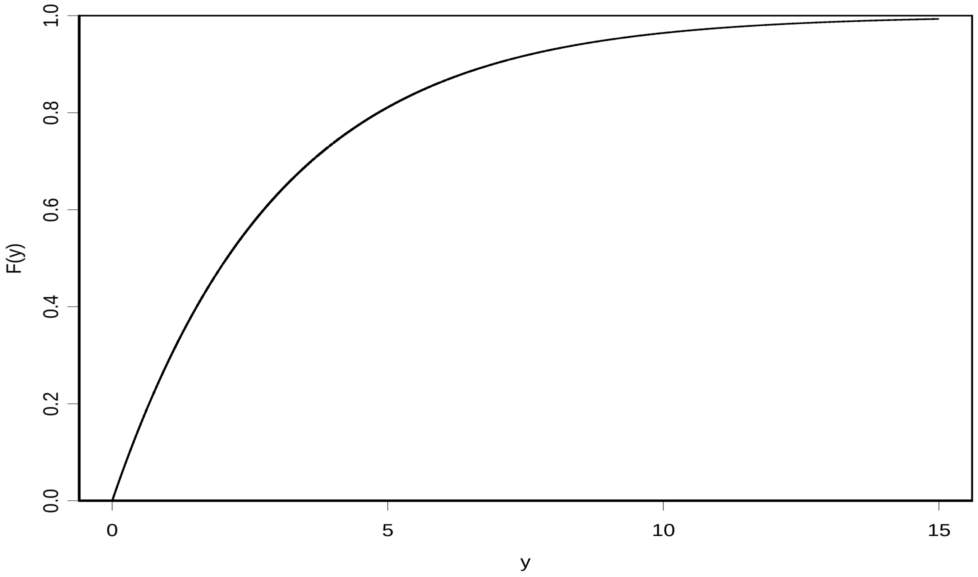

Figure 3.10: *Cumulative distribution function, $F_Y(y)$, in Example* 3.3.

This cdf is depicted in Figure 3.10. Let's calculate the pdf of Y. Again, we need to consider all possible cases:

- when $y < 0$,
$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}0 = 0;$$
- when $y \geq 0$,
$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\left(1 - e^{-y/3}\right) = \frac{1}{3}e^{-y/3}.$$

Thus, putting it all together we get

$$f_Y(y) = \begin{cases} \frac{1}{3}e^{-y/3}, & y \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

This pdf is depicted in Figure 3.11. □

Exercise: For the cdfs in Examples 3.2 and 3.3, verify that these functions satisfy the four theoretical properties for any cdf.

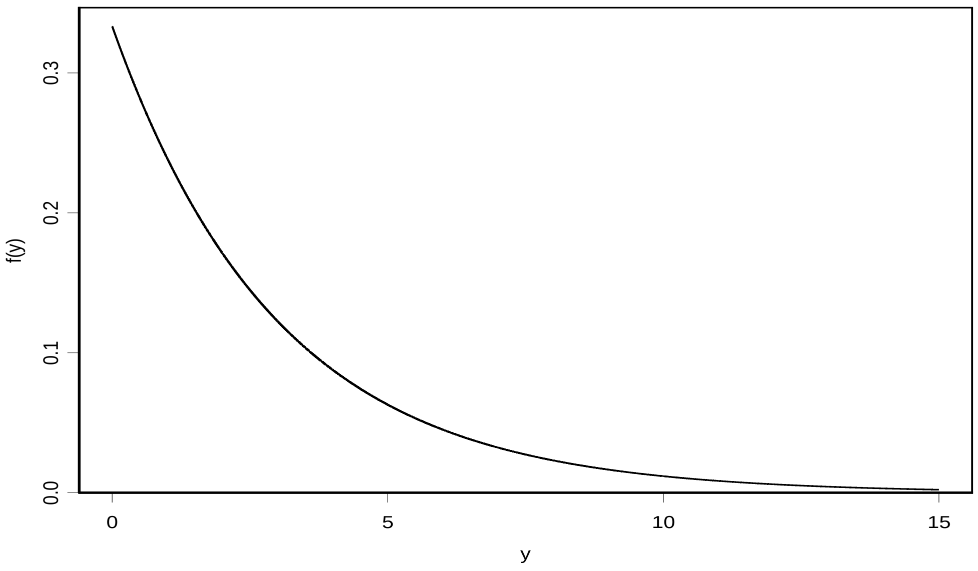

Figure 3.11: *Probability density function, $f_Y(y)$, in Example* 3.3. *This is a probability model for leukemia remission times.*

UBIQUITOUS RESULT: Recall that one of the properties of a continuous pdf is that

$$P(Y \in B) = \int_B f_Y(y)dy,$$

for any $B \subset \mathcal{R}$. If $B = \{y : a \leq y \leq b\}$; i.e., $B = [a, b]$, then

$$P(a \leq Y \leq b) = \int_a^b f_Y(y)dy = F_Y(b) - F_Y(a).$$

Example 3.4. In Example 3.3, what is the probability that a randomly selected patient will have a remission time between 2 and 5 months? That is, what is $P(2 \leq Y \leq 5)$?

SOLUTION. We can attack this **two** ways: one using the cdf, one with the pdf.

- **CDF** (refer to Figure 3.10).

$$\begin{aligned} P(2 \leq Y \leq 5) &= F_Y(5) - F_Y(2) \\ &= (1 - e^{-5/3}) - (1 - e^{-2/3}) \\ &= e^{-2/3} - e^{-5/3} \\ &\approx 0.325. \end{aligned}$$

- **PDF** (refer to Figure 3.11).

$$\begin{aligned} P(2 \leq Y \leq 5) &= \int_2^5 \frac{1}{3} e^{-y/3} dy \\ &= \frac{1}{3} \times (-3) e^{-y/3} \Big|_2^5 \\ &= e^{-2/3} - e^{-5/3} \\ &\approx 0.325. \ \Box \end{aligned}$$

FACT: If Y is a **continuous** random variable with pdf $f_Y(y)$, then $P(Y = a) = 0$ for any real constant a. This follows since

$$P(Y = a) = P(a \leq Y \leq a) = \int_a^a f_Y(y) dy = 0.$$

Thus, for continuous random variables, probabilities are assigned to single points with **zero** probability. This is the key difference between discrete and continuous random variables. An immediate consequence of the above fact is that for any **continuous** random variable Y,

$$P(a \leq Y \leq b) = P(a \leq Y < b) = P(a < Y \leq b) = P(a < Y < b),$$

and the common value is $\int_a^b f_Y(y) dy$.

Example 3.5. Suppose that Y represents the time (in seconds) until a certain chemical reaction takes place (in a manufacturing process, say), and varies according to the pdf

$$f_Y(y) = \begin{cases} cye^{-y/2}, & y \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the c that makes this a valid pdf.

(b) Compute $P(3.5 \leq Y < 4.5)$.

Solution.

(a) To find c, recall that $\int_0^\infty f_Y(y) dy = 1$. Thus,

$$c \int_0^\infty y e^{-y/2} dy = 1.$$

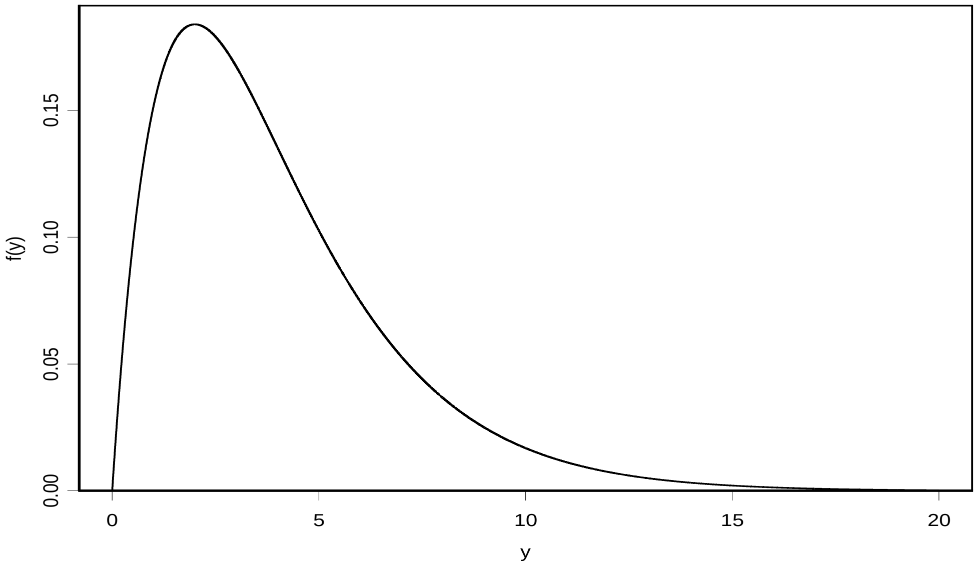

Figure 3.12: *Probability density function, $f_Y(y)$, in Example* 3.5. *This is a probability model for chemical reaction times.*

Using an **integration by parts** argument with $u = y$ and $dv = e^{-y/2}dy$, we have that

$$\begin{aligned} \int_0^\infty ye^{-y/2}dy &= -2ye^{-y/2}\Big|_{y=0}^{\infty} + \int_{y=0}^{\infty} 2e^{-y/2}dy \\ &= (0+0) + \left[2(-2)e^{-y/2}\Big|_{y=0}^{\infty}\right] \\ &= (-4)(0-1) = 4. \end{aligned}$$

Solving for c, we get $c = 1/4$. This pdf is depicted in Figure 3.12.

(b) Using integration by parts again, we get

$$P(3.5 \leq Y < 4.5) = \int_{3.5}^{4.5} \frac{1}{4}ye^{-y/2}dy \approx 0.135.$$

Thus, the probability that the chemical reaction takes place between 3.5 and 4.5 seconds is about 0.14. □

DISCLAIMER: We will use integration by parts repeatedly in this course!!

3.4 Mathematical expectation

3.4.1 Expected values

TERMINOLOGY: Let Y be a continuous random variable with pdf $f_Y(y)$ and support R. The **expected value** (or **mean**) of Y is given by

$$E(Y) = \int_R y f_Y(y) dy.$$

If $E(Y) = \infty$, we say that the expected value does not exist.

RECALL: When Y is a **discrete** random variable with pmf $p_Y(y)$, the expected value of Y is

$$E(Y) = \sum_{y \in R} y p_Y(y).$$

So again, we have the obvious similarities between the continuous and discrete cases.

Example 3.6. Suppose that Y has a pdf given by

$$f_Y(y) = \begin{cases} 2y, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

This pdf is depicted in Figure 3.13. Here, the expected value of Y is given by

$$\begin{aligned} E(Y) &= \int_0^1 y f_Y(y) dy \\ &= \int_0^1 2y^2 dy \\ &= 2\left(\frac{y^3}{3}\bigg|_0^1\right) = 2\left(\frac{1}{3} - 0\right) = 2/3. \quad \Box \end{aligned}$$

EXPECTATIONS OF FUNCTIONS OF Y: Let Y be a continuous random variable with pdf $f_Y(y)$ and support R, and suppose that g is a real-valued function. Then, $g(Y)$ is a random variable and

$$E[g(Y)] = \int_R g(y) f_Y(y) dy.$$

If $E[g(Y)] = \infty$, we say that the expected value does not exist.

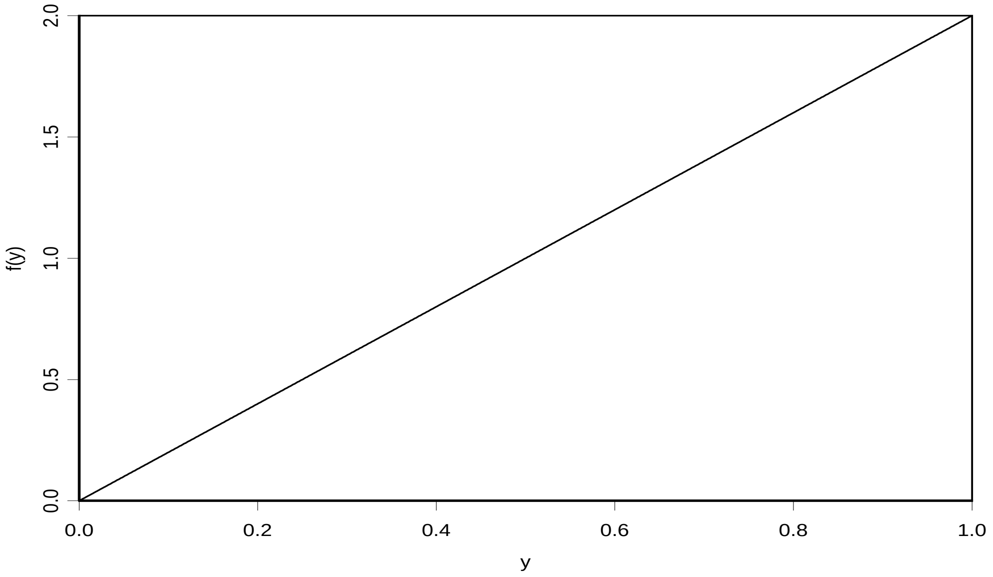

Figure 3.13: *Probability density function, $f_Y(y)$, in Example* 3.6.

Example 3.7. With the pdf in Example 3.6, compute $E(Y^2)$ and $E(\ln Y)$.

SOLUTIONS.

$$E(Y^2) = \int_0^1 2y^3 dy = 2\left(\frac{y^4}{4}\bigg|_0^1\right) = 1/2.$$

Using integration by parts, with $u = \ln y$ and $dv = ydy$,

$$E(\ln Y) = 2\int_0^1 y \ln y dy = 2\left(\frac{1}{2}y^2 \ln y\bigg|_0^1 - \int_0^1 \frac{1}{2}y^2 \times \frac{1}{y}dy\right) = -2\left[\frac{1}{2}\left(\frac{y^2}{2}\bigg|_0^1\right)\right] = -\frac{1}{2}. \ \square$$

PROPERTIES OF EYPECTATIONS: Let Y be a continuous random variable with pdf $f_Y(y)$ and support R, suppose that $g, g_1, g_2, ..., g_k$ are real-valued functions, and let c be any real constant. Then

(a) $E(c) = c$

(b) $E[cg(Y)] = cE[g(Y)]$

(c) $E[\sum_{j=1}^k g_j(Y)] = \sum_{j=1}^k E[g_j(Y)]$.

These properties are identical to those we discussed in the discrete case.

3.4.2 Variance

A SPECIAL EXPECTATION: Let Y be a continuous random variable with pdf $f_Y(y)$, support R, and mean μ. The **variance** of Y is given by

$$\sigma^2 \equiv V(Y) \equiv E[(Y-\mu)^2] = \int_R (y-\mu)^2 f_Y(y)dy.$$

Example 3.8. With the pdf in Example 3.6,

$$f_Y(y) = \begin{cases} 2y, & 0 < y < 1 \\ 0, & \text{otherwise,} \end{cases}$$

compute $\sigma^2 = V(Y)$.

Solutions. Recall that $\mu = E(Y) = 2/3$, from Example 3.6. Using the definition above, the variance of Y is

$$\sigma^2 = V(Y) = \int_0^1 \left(y - \frac{2}{3}\right)^2 \times 2ydy = \frac{1}{18}.$$

Alternatively, we could use the **variance computing formula**; i.e.,

$$V(Y) = E(Y^2) - [E(Y)]^2.$$

We know $E(Y) = 2/3$ and $E(Y^2) = 1/2$ (from Example 3.7). Thus,

$$\sigma^2 = V(Y) = (1/2) - (2/3)^2 = 1/18. \ \square$$

3.4.3 Moment generating functions

ANOTHER SPECIAL EXPECTATION: Let Y be a continuous random variable with pdf $f_Y(y)$ and support R. The **moment generating function** (**mgf**) for Y, denoted by $m_Y(t)$, is given by

$$m_Y(t) = E(e^{tY}) = \int_R e^{ty} f_Y(y),$$

provided $E(e^{tY}) < \infty$ for t in an open neighborhood about 0; i.e., there exists some $h > 0$ such that $E(e^{tY}) < \infty$ for all $t \in (-h, h)$. If $E(e^{tY})$ does not exist in an open neighborhood of 0, we say that the moment generating function does not exist.

Example 3.9. Suppose that Y has a pdf given by

$$f_Y(y) = \begin{cases} e^{-y}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Find the moment generating function of Y and use it to compute $E(Y)$ and $V(Y)$.

Solution.

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \int_0^\infty e^{ty} f_Y(y) dy \\ &= \int_0^\infty e^{-y(1-t)} dy \\ &= -\left(\frac{1}{1-t}\right) e^{-y(1-t)} \Big|_{y=0}^{\infty} \\ &= \left(\frac{1}{1-t}\right), \end{aligned}$$

for values of $t < 1$. With the mgf, we can calculate the mean and variance. Differentiating the mgf, we get

$$M'_Y(t) \equiv \frac{d}{dt} m_Y(t) = \frac{d}{dt}\left(\frac{1}{1-t}\right) = \left(\frac{1}{1-t}\right)^2.$$

Thus,

$$E(Y) = \frac{d}{dt} m_Y(t)\Big|_{t=0} = \left(\frac{1}{1-0}\right)^2 = 1.$$

To find the variance, we first find the second moment:

$$\frac{d^2}{dt^2} m_Y(t) = \frac{d}{dt} \underbrace{\left(\frac{1}{1-t}\right)^2}_{M'_Y(t)} = 2\left(\frac{1}{1-t}\right)^3.$$

Thus, the second moment is

$$E(Y^2) = 2\left(\frac{1}{1-0}\right)^3 = 2.$$

The computing formula gives

$$\sigma^2 = V(Y) = E(Y^2) - [E(Y)]^2 = 2 - 1^2 = 1. \ \square$$

Exercise. Find $E(Y)$ and $V(Y)$ directly (i.e., do not use the mgf). Are your answers the same as above?

3.5 Uniform distribution

TERMINOLOGY: A random variable Y is said to have a **uniform distribution** from θ_1 to θ_2 $(\theta_1 < \theta_2)$ if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{1}{\theta_2-\theta_1}, & \theta_1 < y < \theta_2 \\ 0, & \text{otherwise.} \end{cases}$$

Shorthand notation is $Y \sim \mathcal{U}(\theta_1, \theta_2)$. That the $\mathcal{U}(\theta_1, \theta_2)$ pdf integrates to one is obvious since

$$\int_{\theta_1}^{\theta_2} \frac{1}{\theta_2 - \theta_1}\, dy = \frac{y}{\theta_2 - \theta_1}\bigg|_{\theta_1}^{\theta_2} = \frac{\theta_2 - \theta_1}{\theta_2 - \theta_1} = 1.$$

REMARKS: Sometimes, we call θ_1 and θ_2 the **model parameters**. A popular member of the $\mathcal{U}(\theta_1, \theta_2)$ family is the $\mathcal{U}(0, 1)$ distribution; i.e., a uniform distribution with $\theta_1 = 0$ and $\theta_2 = 1$; this model is used extensively in computer programs to simulate random numbers. The pdf for a $\mathcal{U}(0, 2)$ random variable is depicted in Figure 3.9.

UNIFORM CDF: The cdf $F_Y(y)$ for a $\mathcal{U}(\theta_1, \theta_2)$ distribution is given by

$$F_Y(y) = \begin{cases} 0, & y \leq \theta_1 \\ \frac{y-\theta_1}{\theta_2-\theta_1}, & \theta_1 < y < \theta_2 \\ 1, & y \geq \theta_2. \end{cases}$$

Example 3.10. In a sedimentation experiment, the size of particles studied are uniformly distributed between 0.1 and 0.5 millimeters. What proportion of particles are less than 0.4 millimeters?

Solution. Let Y denote the size of a randomly selected particle. Then, $Y \sim \mathcal{U}(0.1, 0.5)$ and

$$P(Y < 0.4) = \int_{0.1}^{0.4} \frac{1}{0.5 - 0.1} dy = \frac{y}{0.4}\bigg|_{0.1}^{0.4} = \frac{0.3}{0.4} = 0.75. \ \Box$$

MEAN AND VARIANCE: If $Y \sim \mathcal{U}(\theta_1, \theta_2)$, then

$$E(Y) = \frac{\theta_1 + \theta_2}{2}$$

and

$$V(Y) = \frac{(\theta_2 - \theta_1)^2}{12}.$$

These values can be computed using the pdf directly (try it!) or by using the mgf below.

MOMENT GENERATING FUNCTION: Suppose that $Y \sim \mathcal{U}(\theta_1, \theta_2)$. The mgf of Y is given by

$$m_Y(t) = \begin{cases} \frac{e^{\theta_2 t} - e^{\theta_1 t}}{t(\theta_2 - \theta_1)}, & t \neq 0 \\ 1, & t = 0 \end{cases}$$

3.6 Normal distribution

TERMINOLOGY: A random variable Y is said to have a **normal distribution** if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Shorthand notation is $Y \sim \mathcal{N}(\mu, \sigma^2)$. There are two **parameters** in the normal distribution: the **mean** μ and the **variance** σ^2.

FACTS ABOUT ANY NORMAL DISTRIBUTION:

(a) The pdf is **symmetric** about μ; that is, for any $a \in \mathcal{R}$, $f_Y(\mu - a) = f_Y(\mu + a)$.

(b) The points of inflection are located at $y = \mu \pm \sigma$.

(c) Any normal distribution can be transformed to a “standard” normal distribution.

(d) $\lim_{y \to \pm\infty} f_Y(y) = 0$.

TERMINOLOGY: A normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ is called the **standard** normal distribution. It is conventional to let Z denote a random variable that follows a standard normal distribution; we often write $Z \sim \mathcal{N}(0, 1)$.

IMPORTANT: Tabled values of the **standard normal probabilities** are given in Appendix III (Table 4, pp 792) of WMS. This table turns out to be very helpful since the integral

$$F_Y(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

does not exist in closed form! Specifically, the table provides values of

$$1 - F_Z(z) = P(Z > z) = \int_{z}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-u^2/2} du.$$

As mentioned, any normal distribution can be transformed to a "standard" normal distribution (we'll see how later), so there is only a need for one table of probabilities. Of course, probabilities like $P(Z > z)$ can be obtained using software too.

VALIDITY: To show that the $\mathcal{N}(\mu, \sigma^2)$ pdf integrates to one, let $z = \frac{y-\mu}{\sigma}$. Then, $dz = \frac{1}{\sigma}dy$ and $dy = \sigma dz$. Now, define

$$\begin{aligned} I &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz. \end{aligned}$$

Since $I > 0$, it suffices to show that $I^2 = 1$. However, note that

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \times \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left[-\left(\frac{x^2+y^2}{2}\right)\right] dxdy. \end{aligned}$$

Now, switching to polar coordinates; i.e., letting $x = r\cos\theta$ and $y = r\sin\theta$, we get $x^2 + y^2 = r^2(\cos^2\theta + \sin^2\theta) = r^2$, and $dxdy = rdrd\theta$; i.e., the **Jacobian** of the transformation from (x, y) space to (r, θ) space. Thus, we write

$$\begin{aligned} I^2 &= \int_{\theta=0}^{2\pi} \int_{r=0}^{\infty} \frac{1}{2\pi} e^{-r^2/2} rdrd\theta \\ &= \frac{1}{2\pi} \int_{\theta=0}^{2\pi} \left[\int_{r=0}^{\infty} re^{-r^2/2} dr \right] d\theta \\ &= \frac{1}{2\pi} \int_{\theta=0}^{2\pi} \left[- e^{-r^2/2} \Big|_{r=0}^{\infty} \right] d\theta \\ &= \frac{1}{2\pi} \int_{\theta=0}^{2\pi} d\theta = \frac{\theta}{2\pi} \Big|_{\theta=0}^{2\pi} = 1. \ \square \end{aligned}$$

MOMENT GENERATING FUNCTION: Suppose that $Y \sim \mathcal{N}(\mu, \sigma^2)$. The mgf of Y, defined for all t, is given by

$$m_Y(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

Proof.

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \int_{-\infty}^{\infty} e^{ty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{ty-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy. \end{aligned}$$

Define $b = ty - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2$, the exponent in the last integral. Then,

$$\begin{aligned} b &= ty - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2 \\ &= ty - \frac{1}{2\sigma^2}(y^2 - 2\mu y + \mu^2) \\ &= -\frac{1}{2\sigma^2}(y^2 - 2\mu y - 2\sigma^2 ty + \mu^2) \\ &= -\frac{1}{2\sigma^2}[\underbrace{y^2 - 2(\mu+\sigma^2 t)y}_{\text{complete the square}} + \mu^2] \\ &= -\frac{1}{2\sigma^2}\left[y^2 - 2(\mu+\sigma^2 t)y + \underbrace{(\mu+\sigma^2 t)^2 - (\mu+\sigma^2 t)^2}_{\text{add and subtract}} + \mu^2\right] \\ &= -\frac{1}{2\sigma^2}\left\{[y - (\mu+\sigma^2 t)]^2\right\} + \frac{1}{2\sigma^2}\left[(\mu+\sigma^2 t)^2 - \mu^2\right] \\ &= -\frac{1}{2\sigma^2}(y-a)^2 + \underbrace{\frac{1}{2\sigma^2}(\mu^2 + 2\mu\sigma^2 t + \sigma^4 t^2 - \mu^2)}_{=c,\text{ say}}, \end{aligned}$$

where $a = \mu + \sigma^2 t$. Thus, the last integral above is equal to

$$\left(\int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-a)^2}}_{\mathcal{N}(a,\sigma^2)\text{ density}} dy\right) \times e^c$$

Now, finally note $e^c \equiv \exp(c) = \exp(\mu t + \sigma^2 t^2/2)$. Thus, the result follows. □

Exercise: Use the mgf to verify that $E(Y) = \mu$ and $V(Y) = \sigma^2$.

IMPORTANT: Suppose that $Y \sim \mathcal{N}(\mu, \sigma^2)$. Then, the random variable

$$Z = \frac{Y-\mu}{\sigma}$$

has a normal distribution with mean 0 and variance 1. That is, $Z \sim \mathcal{N}(0, 1)$.

Proof. Let $Z = \frac{1}{\sigma}(Y - \mu)$. The mgf of Z is given by

$$\begin{aligned} m_Z(t) = E(e^{tZ}) &= E[\exp(tZ)] \\ &= E\left\{\exp\left[t\left(\frac{Y-\mu}{\sigma}\right)\right]\right\} \\ &= E\left[\exp(-\mu t/\sigma)\exp\left(\frac{t}{\sigma}Y\right)\right] \\ &= \exp(-\mu t/\sigma)\underbrace{E\left[\exp\left(\frac{t}{\sigma}Y\right)\right]}_{m_Y(t/\sigma)} \\ &= \exp(-\mu t/\sigma)\times\exp\left[\mu(t/\sigma) + \frac{\sigma^2(t/\sigma)^2}{2}\right] = e^{t^2/2}, \end{aligned}$$

which is the mgf of a $\mathcal{N}(0,1)$ random variable. Thus, by the **uniqueness** of moment generating functions, we know that $Z \sim \mathcal{N}(0,1)$. □

USEFULNESS: From the last result, we know that if $Y \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\{y_1 < Y < y_2\} = \left\{\frac{y_1-\mu}{\sigma} < \frac{Y-\mu}{\sigma} < \frac{y_2-\mu}{\sigma}\right\} = \left\{\frac{y_1-\mu}{\sigma} < Z < \frac{y_2-\mu}{\sigma}\right\}.$$

As a result,

$$\begin{aligned} P(y_1 < Y < y_2) &= P\left(\frac{y_1-\mu}{\sigma} < Z < \frac{y_2-\mu}{\sigma}\right) \\ &\equiv \Phi\left(\frac{y_2-\mu}{\sigma}\right) - \Phi\left(\frac{y_1-\mu}{\sigma}\right), \end{aligned}$$

where $\Phi(\cdot)$ denotes the **cdf** of the $\mathcal{N}(0,1)$ distribution. Note also that $\Phi(-z) = 1 - \Phi(z)$, for $z > 0$.

Example 3.11. In Florida, young large-mouth bass were studied to examine the level of mercury contamination, Y (measured in parts per million), which varies according to a normal distribution with mean $\mu = 18$ and variance $\sigma^2 = 16$. This model is depicted in Figure 3.14.

(a) What proportion of contamination levels are between 11 and 21 parts per million?

(b) For this model, ninety percent of all contamination levels will be above what mercury level?

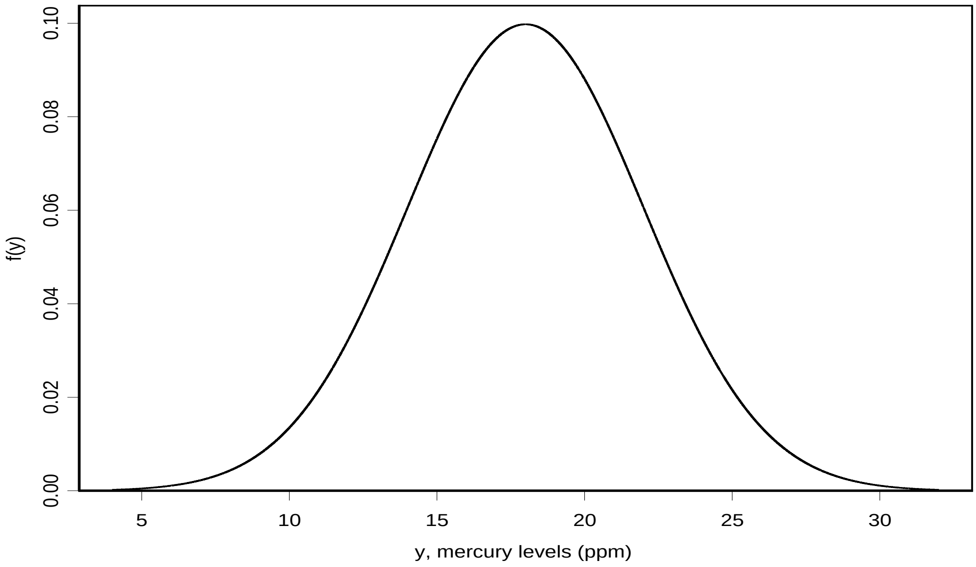

Figure 3.14: *Probability density function, $f_Y(y)$, in Example* 3.11*. A model for mercury contamination in large-mouth bass.*

Solutions: (a) In this part, we want $P(11 < Y < 21)$. By standardizing, we see that

$$\begin{aligned} P(11 < Y < 21) &= P\left(\frac{11-18}{4} < \frac{Y-18}{4} < \frac{21-18}{4}\right) \\ &= P\left(\frac{11-18}{4} < Z < \frac{21-18}{4}\right) \\ &= P(-1.75 < Z < 0.75) \\ &= \Phi(0.75) - \Phi(-1.75) = 0.7734 - 0.0401 = 0.7333. \end{aligned}$$

For (b), we want to find the **10th percentile** of the $Y \sim \mathcal{N}(18, 16)$ distribution; i.e., we want the value y such that $0.90 = P(Y > y) = 1 - F_Y(y)$. To find y, first we'll find the z so that $0.90 = P(Z > z) = 1 - \Phi(z)$, then we'll "unstandardize" y. From Table 4, we see $z = -1.28$ so that

$$\frac{y-18}{4} = -1.28 \implies y = 12.88.$$

Thus, 90 percent of all contamination levels are larger that 12.88 parts per million. □

3.7 The gamma family of pdfs

THE GAMMA FAMILY: In this section, we examine an important family of probability distributions; namely, those in the **gamma family**. There are three "named distributions" in particular:

- exponential distribution
- gamma distribution
- χ^2 distribution

NOTE: The exponential and gamma distributions are popular models for **lifetime** random variables; i.e., random variables that record "time to event" measurements, such as the lifetimes of an electrical component, death times for human subjects, etc. Other **lifetime distributions** include the lognormal, Weibull, and loggamma probability models.

3.7.1 Exponential distribution

TERMINOLOGY: A random variable Y is said to have an **exponential distribution** with parameter $\beta > 0$ if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{1}{\beta}e^{-y/\beta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

NOTATION: Shorthand notation is $Y \sim \text{exponential}(\beta)$. The value β determines the scale of the distribution (it is sometimes called the **scale parameter**). That the exponential density function integrates to one is easily shown (verify!).

MOMENT GENERATING FUNCTION: Suppose that $Y \sim \text{exponential}(\beta)$. The mgf of Y is given by

$$m_Y(t) = \frac{1}{1-\beta t},$$

for values of $t < 1/\beta$.

Proof. Let $\beta = \eta(1+\eta t)^{-1}$ so that $\eta = \beta(1-\beta t)^{-1}$ and $ty - y/\beta = -y/\eta$. Then,

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \int_0^\infty e^{ty}\frac{1}{\beta}e^{-y/\beta}dy \\ &= \frac{1}{\beta}\int_0^\infty e^{-y/\eta}dy \\ &= -\frac{\eta}{\beta}e^{-y/\eta}\Big|_{y=0}^{\infty} \\ &= \frac{1}{1-\beta t}. \end{aligned}$$

Note that for the last expression to be correct, we need $\eta > 0$; i.e., we need $t < \frac{1}{\beta}$. □

MEAN AND VARIANCE: Suppose that $Y \sim$ exponential(β). The mean and variance of Y are given by

$$E(Y) = \beta$$

and

$$V(Y) = \beta^2.$$

Proof: Exercise. □

Example 3.12. The lifetime of a certain electrical component has an exponential distribution with mean $\beta = 500$ hours. Engineers using this component are particularly interested in the probability until failure. What is the probability that a randomly selected component fails before 100 hours? lasts between 250 and 750 hours?

Solution. With $\beta = 500$, the pdf for Y is given by

$$f_Y(y) = \begin{cases} \frac{1}{500}e^{-y/500}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

This pdf is depicted in Figure 3.15. Thus, the probability of failing before 100 hours is given by

$$P(Y < 100) = \int_0^{100} \frac{1}{500}e^{-y/500}dy \approx 0.181.$$

Similarly, the probability of failing between 250 and 750 hours is

$$P(250 < Y < 750) = \int_{250}^{750} \frac{1}{500}e^{-y/500}dy \approx 0.383. \quad \square$$

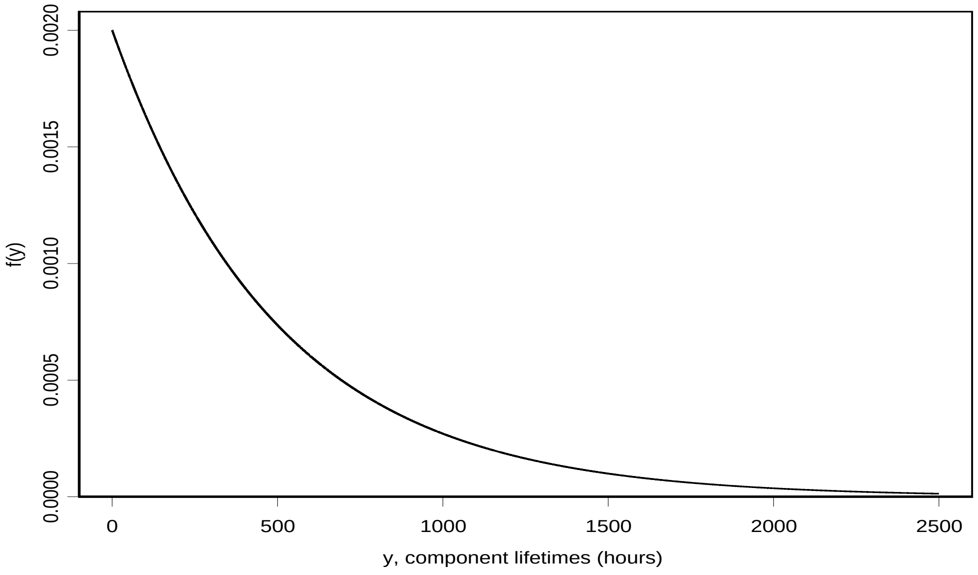

Figure 3.15: *Probability density function, $f_Y(y)$, in Example* 3.12. *A model for electrical component lifetimes.*

CUMULATIVE DISTRIBUTION FUNCTION: Suppose that $Y \sim$ exponential(β). Then, the cdf of Y exists in closed form and is given by

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ 1 - e^{-y/\beta}, & y > 0. \end{cases}$$

The cdf for the exponential random variable in Example 3.12 is depicted in Figure 3.16.

THE MEMORYLESS PROPERTY: Suppose that $Y \sim$ exponential(β), and suppose that r and s are both positive constants. Then

$$P(Y > r + s | Y > r) = P(Y > s).$$

That is, given that the lifetime Y has exceeded r, the probability that Y exceeds $r+s$ (i.e., an additional s units) is the same as if we were to look at Y unconditionally lasting until time s. Put another way, that Y has actually "made it" to time r has been forgotten! The exponential random variable is the only continuous random variable that enjoys the memoryless property.

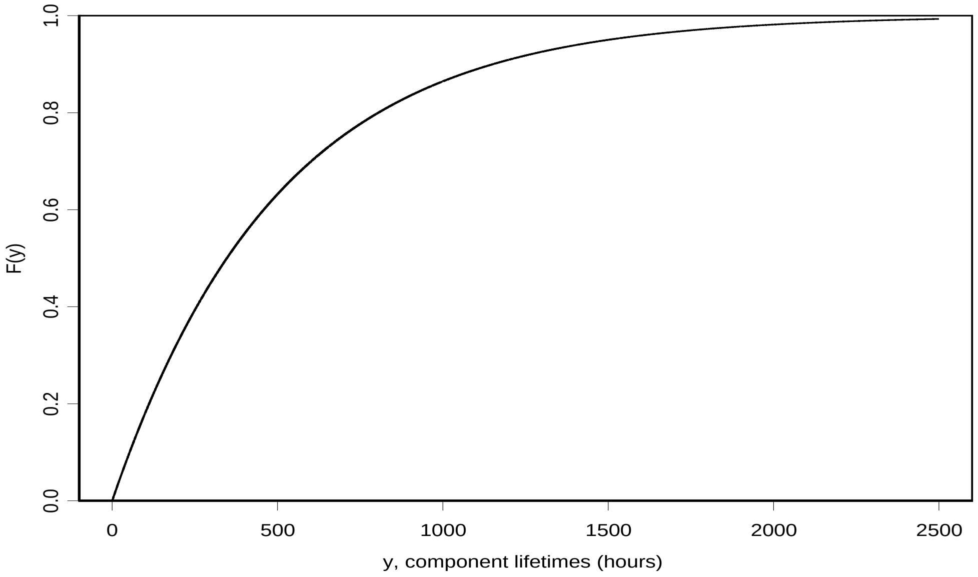

Figure 3.16: *Cumulative distribution function, $F_Y(y)$, in Example* 3.12*. A model for electrical component lifetimes.*

RELATIONSHIP WITH A POISSON PROCESS: Suppose that we are observing events according to a Poisson process with rate $\lambda = 1/\beta$, and let the random variable W denote the time until the first occurrence. Then, $W \sim \text{exponential}(\beta)$.

Proof: Clearly, W is a continuous random variable with nonnegative support. Thus, for $w \geq 0$, we have

$$\begin{aligned} F_W(w) = P(W \leq w) &= 1 - P(W > w) \\ &= 1 - P(\{\text{no events in } [0, w]\}) \\ &= 1 - \frac{e^{-\lambda w}(\lambda w)^0}{0!} \\ &= 1 - e^{-\lambda w}. \end{aligned}$$

Substituting $\lambda = 1/\beta$, we find that

$$F_W(w) = 1 - e^{-w/\beta},$$

the cdf of an exponential random variable with mean β. Thus, the result follows. $\square$

3.7.2 Gamma distribution

THE GAMMA FUNCTION: The **gamma function** is a function of t, defined for all $t > 0$ as

$$\Gamma(t) = \int_0^\infty y^{t-1}e^{-y}dy$$

FACTS ABOUT THE GAMMA FUNCTION:

(1) A simple argument shows that $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$, for all $\alpha > 1$.

(2) If α is an integer, $\Gamma(\alpha) = (\alpha - 1)!$. For example, $\Gamma(5) = 4! = 24$.

TERMINOLOGY: A random variable Y is said to have a **gamma distribution** with parameters $\alpha > 0$ and $\beta > 0$ if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha}y^{\alpha-1}e^{-y/\beta}, & y > 0 \\ 0, & \text{otherwise}. \end{cases}$$

Shorthand notation is $Y \sim \text{gamma}(\alpha, \beta)$.

REMARK: This model is indexed by two parameters. We call α the **shape parameter** and β the **scale parameter**. The gamma probability model is extremely flexible! By changing the values of α and β, the gamma pdf can assume many shapes. Thus, the gamma model is very popular for modeling lifetime data.

IMPORTANT NOTE: When $\alpha = 1$, the gamma pdf reduces to the exponential(β) pdf!

REMARK: To see that the gamma pdf integrates to one, consider the change of variable $u = y/\beta$. Then, $du = \frac{1}{\beta}dy$ and

$$\int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha}y^{\alpha-1}e^{-y/\beta}dy = \frac{1}{\Gamma(\alpha)}\int_0^\infty u^{\alpha-1}e^{-u}du = \frac{\Gamma(\alpha)}{\Gamma(\alpha)} = 1. \ \Box$$

MGF FOR THE GAMMA DISTRIBUTION: Suppose that $Y \sim \text{gamma}(\alpha, \beta)$. Then, for values of $t < 1/\beta$, the mgf of Y is given by

$$m_Y(t) = \left(\frac{1}{1 - \beta t}\right)^\alpha.$$

Proof. Let $\beta = \eta(1+\eta t)^{-1}$ so that $\eta = \beta(1-\beta t)^{-1}$ and $ty - y/\beta = -y/\eta$.

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \int_0^\infty e^{ty}\frac{1}{\Gamma(\alpha)\beta^\alpha}y^{\alpha-1}e^{-y/\beta}dy \\ &= \frac{1}{\beta^\alpha}\int_0^\infty \frac{1}{\Gamma(\alpha)}y^{\alpha-1}e^{-y/\eta}dy \\ &= \frac{\eta^\alpha}{\beta^\alpha}\int_0^\infty \underbrace{\frac{1}{\Gamma(\alpha)\eta^\alpha}y^{\alpha-1}e^{-y/\eta}}_{\text{gamma}(\alpha,\eta)\text{ density}}dy \\ &= \left(\frac{\eta}{\beta}\right)^\alpha = \left(\frac{1}{1-\beta t}\right)^\alpha. \ \square \end{aligned}$$

MEAN AND VARIANCE: If $Y \sim \text{gamma}(\alpha,\beta)$, then

$$E(Y) = \alpha\beta \quad \text{and} \quad V(Y) = \alpha\beta^2.$$

Proof. Exercise. □

TERMINOLOGY: When talking about the gamma(α,β) density function, it is often helpful to think of the formula in two parts:

- the **kernel**: $y^{\alpha-1}e^{-y/\beta}$
- the **constant**: $[\Gamma(\alpha)\beta^\alpha]^{-1}$

Example 3.13. Suppose that Y has pdf given by

$$f_Y(y) = \begin{cases} cy^2e^{-y/4}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(a) What is the value of c that makes this a valid pdf?

(b) Give an integral expression that equals $P(Y < 8)$? How could we solve this equation?

(c) What is the mgf of Y?

(d) What is the mean and standard deviation of Y?

RELATIONSHIP WITH A POISSON PROCESS: Suppose that we are observing events according to a Poisson process with rate $\lambda = 1/\beta$, and let the random variable W denote the time until the αth occurrence. Then, $W \sim \text{gamma}(\alpha,\beta)$.

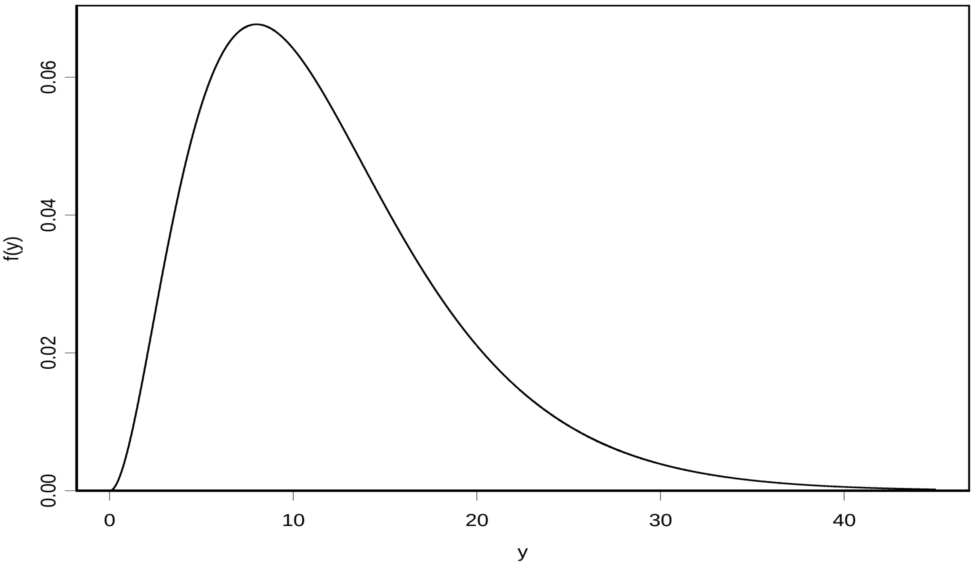

Figure 3.17: *Probability density function, $f_Y(y)$, in Example* 3.13.

Proof: Clearly, W is continuous with nonnegative support. Thus, for $w \geq 0$, we have

$$\begin{aligned} F_W(w) = P(W \leq w) &= 1 - P(W > w) \\ &= 1 - P(\{\text{fewer than } \alpha \text{ events in } [0, w]\}) \\ &= 1 - \sum_{j=0}^{\alpha-1} \frac{e^{-\lambda w}(\lambda w)^j}{j!}. \end{aligned}$$

The pdf of W, $f_W(w)$, is equal to $F'_W(w)$, provided that this derivative exists. For $w > 0$,

$$\begin{aligned} f_W(w) = F'_W(w) &= \lambda e^{-\lambda w} - e^{-\lambda w} \underbrace{\sum_{j=1}^{\alpha-1} \left[\frac{j(\lambda w)^{j-1}\lambda}{j!} - \frac{(\lambda w)^j \lambda}{j!} \right]}_{\text{telescoping sum}} \\ &= \lambda e^{-\lambda w} - e^{-\lambda w} \left[\lambda - \frac{\lambda(\lambda w)^{\alpha-1}}{(\alpha-1)!} \right] \\ &= \frac{\lambda(\lambda w)^{\alpha-1} e^{-\lambda w}}{(\alpha-1)!} = \frac{\lambda^\alpha}{\Gamma(\alpha)} w^{\alpha-1} e^{\lambda w}. \end{aligned}$$

Substituting $\lambda = 1/\beta$,

$$f_W(w) = \frac{1}{\Gamma(\alpha)\beta^\alpha} w^{\alpha-1} e^{-w/\beta},$$

for $w > 0$, which is the pdf for the gamma(α, β) distribution. □

3.7.3 χ^2 distribution

TERMINOLOGY: In the gamma(α, β) family, when $\alpha = \nu/2$, for any integer ν, and $\beta = 2$, we call the resulting distribution a χ^2 **distribution** with ν degrees of freedom. If Y has a χ^2 distribution with ν degrees of freedom, we write $Y \sim \chi^2(\nu)$.

NOTE: At this point, it suffices to know that the χ^2 distribution is really just a "special" gamma distribution. However, it should be noted that the χ^2 distribution is used extensively in applied statistics. Many statistical procedures used in the literature are valid because of this model!

PROBABILITY DENSITY FUNCTION: If $Y \sim \chi^2(\nu)$, then the pdf of Y is given by

$$f_Y(y) = \begin{cases} \frac{1}{\Gamma(\frac{\nu}{2})2^{\nu/2}} y^{(\nu/2)-1} e^{-y/2}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

MOMENT GENERATING FUNCTION: Suppose that $Y \sim \chi^2(\nu)$. Then, for values of $t < 1/2$, the mgf of Y is given by

$$m_Y(t) = \left(\frac{1}{1-2t}\right)^{\nu/2}.$$

Proof. Take the gamma(α, β) mgf and put in $\alpha = \nu/2$ and $\beta = 2$. □

MEAN AND VARIANCE OF THE χ^2 DISTRIBUTION: If $Y \sim \chi^2(\nu)$, then

$$E(Y) = \nu \qquad \text{and} \qquad V(Y) = 2\nu.$$

Proof. Take the gamma(α, β) formulae and substitute $\alpha = \nu/2$ and $\beta = 2$. □

TABLED VALUES FOR CDF: Because the χ^2 distribution is so pervasive in applied statistics, tables of probabilities are common. Table 6 (WMS, pp 794-5) provides values of y which satisfy

$$P(Y > y) = \int_y^\infty \frac{1}{\Gamma(\frac{\nu}{2})2^{\nu/2}} u^{(\nu/2)-1} e^{-u/2} du$$

for different values of y and degrees of freedom ν.

3.8 Beta distribution

TERMINOLOGY: A random variable Y is said to have a **beta distribution** with parameters $\alpha > 0$ and $\beta > 0$ if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Since the support of Y is $0 < y < 1$, the beta distribution is a popular probability model for **proportions**. Shorthand notation is $Y \sim \text{beta}(\alpha, \beta)$. The constant $B(\alpha, \beta)$ is given by

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

TERMINOLOGY: When talking about the beta(α, β) density function, it is often helpful to think of the formula in two parts:

- the **kernel**: $y^{\alpha-1}(1-y)^{\beta-1}$
- the **constant**: $\frac{1}{B(\alpha,\beta)}$

THE SHAPE OF THE BETA PDF: The beta pdf is very flexible. That is, by changing the values of α and β, we can come up with many different pdf shapes. See Figure 3.18 for examples.

- When $\alpha = \beta$, the pdf is **symmetric** about the line $y = \frac{1}{2}$.
- When $\alpha < \beta$, the pdf is **skewed right** (i.e., smaller values of y are more likely).
- When $\alpha > \beta$, the pdf is **skewed left** (i.e., larger values of y are more likely).
- When $\alpha = \beta = 1$, the beta pdf reduces to the $\mathcal{U}(0, 1)$ pdf!

MOMENT GENERATING FUNCTION: The mgf of a beta(α, β) random variable exists, but not in a nice compact formula. Hence, we'll compute moments directly.

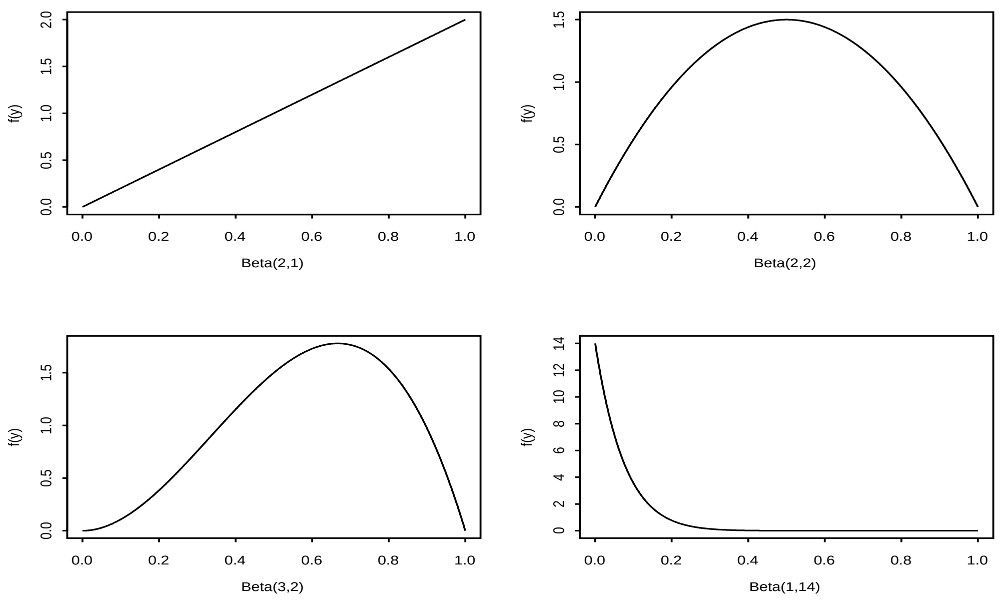

Figure 3.18: *Four different beta probability models.*

MEAN AND VARIANCE OF THE BETA DISTRIBUTION: If $Y \sim \text{beta}(\alpha, \beta)$, then

$$E(Y) = \frac{\alpha}{\alpha+\beta} \qquad \text{and} \qquad V(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Proof. Exercise. □

Example 3.14. A small filling station is supplied with premium gasoline once per day (and can supply at most 1000 gallons). Its daily volume of sales (in 1000s of gallons) is a random variable, say Y, which has the beta distribution

$$f_Y(y) = \begin{cases} 5(1-y)^4, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

(a) What is are the parameters in this distribution; i.e., what are α and β?

(b) What is the **average** daily sales?

(c) What need the capacity of the tank be so that the probability of the supply being exhausted in any day is 0.01?

(d) Treating daily sales as independent (from day to day), what is the probability that during any given 7-day span, there are exactly 2 days where sales exceed 200 gallons?

Solutions. (a) $\alpha = 1$ and $\beta = 5$.

(b) $E(Y) = \frac{1}{1+5} = 1/6$. Thus, the average sales is about 166.66 gallons.

(c) We want to find the capacity, say c, such that $P(Y > c) = 0.01$. This means that

$$P(Y > c) = \int_c^1 5(1-y)^4 dy = 0.01,$$

and we need to solve this equation for c. Using a change of variable $u = 1 - y$,

$$\int_c^1 5(1-y)^4 dy = \int_0^{1-c} 5u^4 dy = u^5 \Big|_0^{1-c} = (1-c)^5.$$

Thus, we have $(1-c)^5 = 0.01 \Rightarrow 1 - c = (0.01)^{1/5} \Rightarrow c = 1 - (0.01)^{1/5} \approx 0.602$, and so there must be about 602 gallons in the tank.

(d) First, we compute

$$P(Y > 0.2) = \int_{0.2}^1 5(1-y)^4 dy = \int_0^{0.8} 5u^4 du = u^5 \Big|_0^{0.8} = (0.8)^5 = 0.328.$$

This is the probability that sales exceed 200 gallons on any given day. Now, **treat each day as a "trial,"** and let X denote the number of days where "sales exceed 200 gallons" (i.e., a "success"). Because days are assumed independent, $X \sim b(7, 0.328)$ and

$$P(X = 2) = \binom{7}{2}(0.328)^2(1-0.328)^5 = 0.310. \ \Box$$

3.9 Chebyshev's Inequality

MARKOV'S INEQUALITY: Suppose that X is a **nonnegative** random variable with pdf (pmf) $f_X(x)$, and let c be any positive constant. Then,

$$P(X > c) \leq \frac{E(X)}{c}.$$

Proof. First, define the event $B = \{x : x > c\}$. We know that

$$\begin{aligned} E(X) = \int_0^\infty x f_X(x)dx &= \int_B x f_X(x)dx + \int_{\overline{B}} x f_X(x)dx \\ &\geq \int_B x f_X(x)dx \\ &\geq \int_B c f_X(x)dx \\ &= cP(X > c). \ \Box \end{aligned}$$

SPECIAL CASE: Let Y be **any random variable**, discrete or continuous, with mean μ and variance $\sigma^2 < \infty$. Then, for $k > 0$,

$$P(|Y - \mu| > k\sigma) \leq \frac{1}{k^2}.$$

This is known as **Chebyshev's Inequality**.

Proof. Apply Markov's Inequality with $X = (Y - \mu)^2$ and $c = k^2\sigma^2$. With these substitutions, we have

$$P(|Y - \mu| > k\sigma) = P[(Y - \mu)^2 > k^2\sigma^2] \leq \frac{E[(Y - \mu)^2]}{k^2\sigma^2} = \frac{1}{k^2}. \ \square$$

REMARK: The beauty of Chebyshev's result is that it applies to **any random variable** Y. In words, $P(|Y - \mu| > k\sigma)$ is the probability that the random variable Y will differ from the mean μ by more than k standard deviations. If we do not know how Y is distributed, we can not compute $P(|Y - \mu| > k\sigma)$ exactly, but, at least we can put an upper bound on this probability; this is what Chebyshev's result allows us to do. Note that

$$P(|Y - \mu| > k\sigma) = 1 - P(|Y - \mu| \leq k\sigma) = 1 - P(\mu - k\sigma \leq Y \leq \mu + k\sigma).$$

Thus, it must be the case that

$$P(|Y - \mu| \leq k\sigma) = P(\mu - k\sigma \leq Y \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}.$$

Example 3.15. Suppose that Y represents the amount of precipitation (in inches) observed annually in Barrow, AK. The exact probability distribution for Y is unknown, but, from historical information, it is posited that $\mu = 4.5$ and $\sigma = 1$. What is a lower bound on the probability that there will be **between** 2.5 and 6.5 inches of precipitation during the next year?

SOLUTION: We want to compute a lower bound for $P(2.5 \leq Y \leq 6.5)$. Note that

$$P(2.5 \leq Y \leq 6.5) = P(|Y - \mu| \leq 2\sigma) \geq 1 - \frac{1}{2^2} = 0.75.$$

Thus, we know that $P(2.5 \leq Y \leq 6.5) \geq 0.75$. The chances are good that, in fact, Y will be between 2.5 and 6.5 inches.

4 Multivariate Distributions

Complementary reading from WMS: Chapter 5.

4.1 Introduction

REMARK: So far, we have only discussed **univariate** (single) random variables (their probability distributions, moment generating functions, means and variances, etc). In practice, however, investigators are often interested in probability statements concerning two or more random variables. Consider the following examples:

- In an agricultural field trial, we might to understand the relationship between yield (Y, measured in bushels/acre) and the nitrogen content of the soil (X).
- In an educational assessment program, we might want to predict a student's posttest score (Y_2) from her pretest score (Y_1).
- In a clinical trial, physicians might want to characterize the concentration of a drug (Y) in one's body as a function of the time (X) from injection.
- In a marketing study, the goal is to forecast next month's sales, say Y_n, based on sales figures from the previous $n-1$ periods, say $Y_1, Y_2, ..., Y_{n-1}$.

GOAL: In each of these examples, our goal is to describe the **relationship** between (or among) the random variables that are recorded. As it turns out, these relationships can be described mathematically through a probabilistic model.

TERMINOLOGY: If Y_1 and Y_2 are random variables, then (Y_1, Y_2) is called a **bivariate random vector**. If $Y_1, Y_2, ..., Y_n$ denote n random variables, then $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ is called an **n-variate random vector**. For much of this chapter, we will consider the $n = 2$ bivariate case. However, all ideas discussed herein extend naturally to higher dimensional settings.

4.2 Discrete random vectors

TERMINOLOGY: Let Y_1 and Y_2 be discrete random variables. Then, (Y_1, Y_2) is called a **discrete random vector**, and the **joint probability mass function (pmf)** of Y_1 and Y_2 is given by

$$p_{Y_1,Y_2}(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2),$$

for all $(y_1, y_2) \in R_{Y_1,Y_2}$. The set $R_{Y_1,Y_2} \subseteq \mathcal{R}^2$ is the two dimensional **support** of (Y_1, Y_2). The function $p_{Y_1,Y_2}(y_1, y_2)$ has the following properties:

(1) $0 \leq p_{Y_1,Y_2}(y_1, y_2) \leq 1$, for all $(y_1, y_2) \in R_{Y_1,Y_2}$

(2) $\sum_{R_{Y_1,Y_2}} p_{Y_1,Y_2}(y_1, y_2) = 1$

(3) $P[(Y_1, Y_2) \in B] = \sum_B p_{Y_1,Y_2}(y_1, y_2)$, for any set $B \subset \mathcal{R}^2$.

Example 4.1. An urn contains 3 red balls, 4 white balls, and 5 green balls. Let (Y_1, Y_2) denote the bivariate random vector where, out of 3 randomly selected balls,

$$\begin{aligned} Y_1 &= \text{number of red balls} \\ Y_2 &= \text{number of white balls.} \end{aligned}$$

Consider the following calculations:

$$\begin{aligned}
p_{Y_1,Y_2}(0,0) = \frac{\binom{3}{0}\binom{4}{0}\binom{5}{3}}{\binom{12}{3}} &= \frac{10}{220} \\
p_{Y_1,Y_2}(0,1) = \frac{\binom{3}{0}\binom{4}{1}\binom{5}{2}}{\binom{12}{3}} &= \frac{40}{220} \\
p_{Y_1,Y_2}(0,2) = \frac{\binom{3}{0}\binom{4}{2}\binom{5}{1}}{\binom{12}{3}} &= \frac{30}{220} \\
p_{Y_1,Y_2}(0,3) = \frac{\binom{3}{0}\binom{4}{3}\binom{5}{0}}{\binom{12}{3}} &= \frac{4}{220} \\
p_{Y_1,Y_2}(1,0) = \frac{\binom{3}{1}\binom{4}{0}\binom{5}{2}}{\binom{12}{3}} &= \frac{30}{220} \\
p_{Y_1,Y_2}(1,1) = \frac{\binom{3}{1}\binom{4}{1}\binom{5}{1}}{\binom{12}{3}} &= \frac{60}{220}
\end{aligned}$$

Table 4.2: *Joint pmf* $p_{Y_1,Y_2}(y_1,y_2)$ *for Example* 4.1 *displayed in tabular form.*

$p_{Y_1,Y_2}(y_1,y_2)$	$y_2=0$	$y_2=1$	$y_2=2$	$y_2=3$
$y_1=0$	$\frac{10}{220}$	$\frac{40}{220}$	$\frac{30}{220}$	$\frac{4}{220}$
$y_1=1$	$\frac{30}{220}$	$\frac{60}{220}$	$\frac{18}{220}$	
$y_1=2$	$\frac{15}{220}$	$\frac{12}{220}$		
$y_1=3$	$\frac{1}{220}$			

and similarly,

$$\begin{aligned}
p_{Y_1,Y_2}(1,2) &= \frac{18}{220}\\
p_{Y_1,Y_2}(2,0) &= \frac{15}{220}\\
p_{Y_1,Y_2}(2,1) &= \frac{12}{220}\\
p_{Y_1,Y_2}(3,0) &= \frac{1}{220}.
\end{aligned}$$

Here, the support is

$$R_{Y_1,Y_2}=\{(0,0),(0,1),(0,2),(0,3),(1,0),(1,1),(1,2),(2,0),(2,1),(3,0)\}.$$

Table 4.2 depicts the joint pmf. It is straightforward to see that $\sum_{R_{Y_1,Y_2}} p_{Y_1,Y_2}(y_1,y_2)=1$.

QUESTION: What is the probability that, among the three balls chosen, there is at most 1 red ball and at most 1 white ball? That is, what is $P(Y_1\leq 1, Y_2\leq 1)$?

SOLUTION. Here, we want to compute $P(B)$, where the set $B=\{(0,0),(0,1),(1,0),(1,1)\}$. From the properties associated with the joint pmf, this calculation is given by

$$\begin{aligned}
P(B)=P(Y_1\leq 1,Y_2\leq 1) &= p_{Y_1,Y_2}(0,0)+p_{Y_1,Y_2}(0,1)+p_{Y_1,Y_2}(1,0)+p_{Y_1,Y_2}(1,1)\\
&= \frac{10}{220}+\frac{40}{220}+\frac{30}{220}+\frac{60}{220}\\
&= \frac{140}{220}.
\end{aligned}$$

QUESTION: What is the probability that, among the three balls chosen, there are at least 2 red balls? That is, what is $P(Y_1\geq 2)$?

4.3 Continuous random vectors

TERMINOLOGY: Let Y_1 and Y_2 be continuous random variables. Then, (Y_1, Y_2) is called a **continuous random vector**, and the **joint probability density function (pdf)** of Y_1 and Y_2 is denoted by $f_{Y_1,Y_2}(y_1, y_2)$. The function $f_{Y_1,Y_2}(y_1, y_2)$ has the following properties:

(1) $f_{Y_1,Y_2}(y_1, y_2) > 0$, for all $(y_1, y_2) \in R_{Y_1,Y_2}$ (the two-dimensional support set)

(2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)dy_1dy_2 = 1$

(3) $P[(Y_1, Y_2) \in B] = \int_B f_{Y_1,Y_2}(y_1, y_2)dy_1dy_2$, for any set $B \subset \mathcal{R}^2$.

REMARK: Of course, we realize that

$$P[(Y_1, Y_2) \in B] = \int_B f_{Y_1,Y_2}(y_1, y_2)dy_1dy_2$$

is really a **double integral** since B is a two-dimensional set in the (y_1, y_2) plane; thus, $P[(Y_1, Y_2) \in B]$ represents the volume under $f_{Y_1,Y_2}(y_1, y_2)$ over B.

TERMINOLOGY: Suppose that (Y_1, Y_2) is a continuous random vector with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$. The **joint cumulative distribution function (cdf)** for (Y_1, Y_2) is given by

$$F_{Y_1,Y_2}(y_1, y_2) \equiv P(Y_1 \leq y_1, Y_2 \leq y_2) = \int_{-\infty}^{y_2}\int_{-\infty}^{y_1} f_{Y_1,Y_2}(r, s)drds,$$

for all $(y_1, y_2) \in \mathcal{R}^2$. It follows upon differentiation that the joint pdf is given by

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{\partial^2}{\partial y_1 \partial y_2} F_{Y_1,Y_2}(y_1, y_2),$$

wherever these mixed partial derivatives are defined.

Example 4.2. Suppose that in a controlled agricultural experiment, we observe the random vector (Y_1, Y_2), where Y_1 = temperature (in Celcius) and Y_2 = precipitation level (in inches), and suppose that the joint pdf of (Y_1, Y_2) is given by

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} cy_1y_2, & 10 < y_1 < 20, 0 < y_2 < 3 \\ 0, & \text{otherwise.} \end{cases}$$

(a) What is the value of c?

(b) Compute $P(Y_1 > 15, Y_2 < 1)$.

(c) Compute $P(Y_2 > Y_1/5)$.

SOLUTIONS: (a) We know that

$$\int_{y_1=10}^{20}\int_{y_2=0}^{3} cy_1y_2 \; dy_2dy_1 = 1$$

since $f_{Y_1,Y_2}(y_1,y_2)$ must integrate to 1 over $R_{Y_1,Y_2} = \{(y_1,y_2) : 10 < y_1 < 20, 0 < y_2 < 3\}$; i.e.,

$$1 = \int_{y_1=10}^{20}\int_{y_2=0}^{3} cy_1y_2 \; dy_2dy_1 = c\int_{y_1=10}^{20} y_1\left(\frac{y_2^2}{2}\Big|_0^3\right)dy_1 = \frac{9c}{2}\left(\frac{y_1^2}{2}\Big|_{10}^{20}\right) = \frac{9c}{2}(150) = 675c.$$

Thus, $c = 1/675$.

(b) Let $B = \{(y_1,y_2) : y_1 > 15, y_2 < 1\}$. The value $P[(Y_1,Y_2) \in B] = P(Y_1 > 15, Y_2 < 1)$ represents the volume under $f_{Y_1,Y_2}(y_1,y_2)$ over the set B; i.e.,

$$\begin{aligned} P[(Y_1,Y_2) \in B] = P(Y_1 > 15, Y_2 < 1) &= \int_{y_1=15}^{20}\int_{y_2=0}^{1} \frac{1}{675} y_1y_2 \; dy_2dy_1 \\ &= \frac{1}{675}\int_{y_1=15}^{20} y_1\left(\frac{y_2^2}{2}\Big|_0^1\right)dy_1 \\ &= \frac{1}{1350}\left(\frac{y_1^2}{2}\Big|_{15}^{20}\right) = \frac{1}{1350}\left(200 - \frac{225}{2}\right) \approx 0.065. \end{aligned}$$

(c) Let $D = \{(y_1,y_2) : y_2 > y_1/5\}$. The quantity $P[(Y_1,Y_2) \in D] = P(Y_2 > Y_1/5)$ represents the volume under $f_{Y_1,Y_2}(y_1,y_2)$ over the set D; i.e.,

$$\begin{aligned} P[(Y_1,Y_2) \in D] = P(Y_2 > Y_1/5) &= \int_{y_2=2}^{3}\int_{y_1=10}^{5y_2} \frac{1}{675} y_1y_2 \; dy_1dy_2 \\ &= \frac{1}{675}\int_{y_2=2}^{3} y_2\left(\frac{y_1^2}{2}\Big|_{10}^{5y_2}\right)dy_2 \\ &= \frac{1}{1350}\int_{y_2=2}^{3}(25y_2^3 - 100y_2)dy_2 \\ &= \frac{1}{1350}\left(\frac{25y_2^4}{4} - 50y_2^2\Big|_2^3\right) \approx 0.116. \end{aligned}$$

NOTE: The key thing to remember that, in parts (b) and (c), the probability is simply the **volume** under the density $f_{Y_1,Y_2}(y_1,y_2)$ over a particular set. It is helpful to draw a picture to get the limits of integration correct!

4.4 Marginal distributions

RECALL: The joint pmf of (Y_1, Y_2) in Example 4.1 is depicted below in Table 4.3. You see that by summing out over the values of y_2 in Table 4.3, we obtain the **row sums**

$$\begin{aligned} P(Y_1 = 0) &= \frac{84}{220} \\ P(Y_1 = 1) &= \frac{108}{220} \\ P(Y_1 = 2) &= \frac{27}{220} \\ P(Y_1 = 3) &= \frac{1}{220} \end{aligned}$$

This represents the **marginal distribution** of Y_1. Similarly, by summing out over the values of y_1, we obtain the **column sums**

$P(Y_2 = 0)$	$P(Y_2 = 1)$	$P(Y_2 = 2)$	$P(Y_2 = 3)$
$\frac{56}{220}$	$\frac{112}{220}$	$\frac{48}{220}$	$\frac{4}{220}$

This represents the **marginal distribution** of Y_2.

Table 4.3: *Joint pmf $p_{Y_1,Y_2}(y_1, y_2)$ displayed in tabular form.*

$p_{Y_1,Y_2}(y_1, y_2)$	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	$y_2 = 3$	Row Sum
$y_1 = 0$	$\frac{10}{220}$	$\frac{40}{220}$	$\frac{30}{220}$	$\frac{4}{220}$	$\frac{84}{220}$
$y_1 = 1$	$\frac{30}{220}$	$\frac{60}{220}$	$\frac{18}{220}$		$\frac{108}{220}$
$y_1 = 2$	$\frac{15}{220}$	$\frac{12}{220}$			$\frac{27}{220}$
$y_1 = 3$	$\frac{1}{220}$				$\frac{1}{220}$
Column sum	$\frac{56}{220}$	$\frac{112}{220}$	$\frac{48}{220}$	$\frac{4}{220}$	1

TERMINOLOGY: Let (Y_1, Y_2) be a **discrete** random vector with pmf $p_{Y_1,Y_2}(y_1, y_2)$. Then the **marginal pmf** of Y_1 is

$$p_{Y_1}(y_1) = \sum_{\text{all } y_2} p_{Y_1,Y_2}(y_1, y_2)$$

and the **marginal pmf** of Y_2 is

$$p_{Y_2}(y_2) = \sum_{\text{all } y_1} p_{Y_1,Y_2}(y_1, y_2).$$

MAIN POINT: In the two-dimensional discrete case, marginal pmfs are obtained by "summing out" over the other variable.

TERMINOLOGY: Let (Y_1, Y_2) be a **continuous** random vector with pdf $f_{Y_1,Y_2}(y_1, y_2)$. Then the **marginal pdf** of Y_1 is

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2) dy_2$$

and the **marginal pdf** of Y_2 is

$$f_{Y_2}(y_2) = \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2) dy_1.$$

MAIN POINT: In the two-dimensional continuous case, marginal pdfs are obtained by "integrating out" over the other variable.

Example 4.3. In a simple genetics model, the proportion, say Y_1, of a population with Trait 1 is always less than the proportion, say Y_2, of a population with trait 2, and the random vector (Y_1, Y_2) has joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 6y_1, & 0 < y_1 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the marginal distributions $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$.

(b) Find the probability that the proportion of individuals with trait 2 exceeds 1/2.

(c) Find the probability that the proportion of individuals with trait 2 is at least twice that of the proportion of individuals with trait 1.

SOLUTIONS: (a) To find the marginal distribution of Y_1, i.e., $f_{Y_1}(y_1)$, we integrate out over y_2. For values of $0 \leq y_1 \leq 1$, we have

$$f_{Y_1}(y_1) = \int_{y_2=y_1}^{1} 6y_1 dy_2 = 6y_1(1 - y_1).$$

Thus, the marginal distribution of Y_1 is given by

$$f_{Y_1}(y_1) = \begin{cases} 6y_1(1 - y_1), & 0 < y_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Of course, we recognize this as a beta distribution with $\alpha = 2$ and $\beta = 2$. That is, marginally, $Y_1 \sim \text{beta}(2, 2)$. To find the marginal distribution of Y_2, i.e., $f_{Y_2}(y_2)$, we integrate out over y_1. For values of $0 \leq y_2 \leq 1$, we have

$$f_{Y_2}(y_2) = \int_{y_1=0}^{y_2} 6y_1 dy_1 = 3y_1^2\Big|_0^{y_2} = 3y_2^2.$$

Thus, the marginal distribution of Y_2 is given by

$$f_{Y_2}(y_2) = \begin{cases} 3y_2^2, & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Of course, we recognize this as a beta distribution with $\alpha = 3$ and $\beta = 1$. That is, marginally, $Y_2 \sim \text{beta}(3, 1)$.

(b) Here, we want to find $P(B)$, where the set $B = \{(y_1, y_2) : 0 < y_1 < y_2, y_2 > 1/2\}$. This probability can be computed two different ways:

(i) using the **joint** distribution $f_{Y_1,Y_2}(y_1, y_2)$ and computing

$$P[(Y_1, Y_2) \in B] = \int_{y_2=0.5}^{1} \int_{y_1=0}^{y_2} 6y_1 \; dy_1 dy_2.$$

(ii) using the **marginal** distribution $f_{Y_2}(y_2)$ and computing

$$P(Y_2 > 1/2) = \int_{y_2=0.5}^{1} 3y_2^2 dy_2.$$

Either way, you will get the same answer! Notice that in (i), you are computing the **volume** under $f_{Y_1,Y_2}(y_1, y_2)$ over the set B. In (ii), you are finding the **area** under $f_{Y_2}(y_2)$ over the set $\{y_2 : y_2 > 1/2\}$.

(c) Here, we want to compute $P(Y_2 \geq 2Y_1)$; i.e., we want to compute $P(D)$, where the set $D = \{(y_1, y_2) : y_2 \geq 2y_1\}$. This equals

$$P[(Y_1, Y_2) \in D] = \int_{y_2=0}^{1} \int_{y_1=0}^{y_2/2} 6y_1 dy_1 dy_2 = 0.25.$$

This is the volume under $f_{Y_1,Y_2}(y_1, y_2)$ over the set D. □

4.5 Conditional distributions

RECALL: For events A and B in a non-empty sample space S, we defined

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

for $P(B) > 0$. Now, suppose that (Y_1, Y_2) is a discrete random vector. If we let $B = \{Y_2 = y_2\}$ and $A = \{Y_1 = y_1\}$, we obtain

$$P(A|B) = \frac{P(Y_1 = y_1, Y_2 = y_2)}{P(Y_2 = y_2)} = \frac{p_{Y_1,Y_2}(y_1, y_2)}{p_{Y_2}(y_2)}.$$

TERMINOLOGY: Suppose that (Y_1, Y_2) is a discrete random vector with joint pmf $p_{Y_1,Y_2}(y_1, y_2)$. We define the **conditional probability mass function (pmf)** of Y_1, given $Y_2 = y_2$, as

$$p_{Y_1|Y_2}(y_1|y_2) = \frac{p_{Y_1,Y_2}(y_1, y_2)}{p_{Y_2}(y_2)},$$

whenever $p_{Y_2}(y_2) > 0$. Similarly, the conditional probability mass function of Y_2, given $Y_1 = y_1$, as

$$p_{Y_2|Y_1}(y_2|y_1) = \frac{p_{Y_1,Y_2}(y_1, y_2)}{p_{Y_1}(y_1)},$$

whenever $p_{Y_1}(y_1) > 0$.

Example 4.4. In Example 4.1, we computed the joint pmf for (Y_1, Y_2). The table below depicts this joint pmf as well as the marginal pmfs.

Table 4.4: *Joint pmf* $p_{Y_1,Y_2}(y_1, y_2)$ *displayed in tabular form.*

$p_{Y_1,Y_2}(y_1, y_2)$	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	$y_2 = 3$	Row Sum
$y_1 = 0$	$\frac{10}{220}$	$\frac{40}{220}$	$\frac{30}{220}$	$\frac{4}{220}$	$\frac{84}{220}$
$y_1 = 1$	$\frac{30}{220}$	$\frac{60}{220}$	$\frac{18}{220}$		$\frac{108}{220}$
$y_1 = 2$	$\frac{15}{220}$	$\frac{12}{220}$			$\frac{27}{220}$
$y_1 = 3$	$\frac{1}{220}$				$\frac{1}{220}$
Column sum	$\frac{56}{220}$	$\frac{112}{220}$	$\frac{48}{220}$	$\frac{4}{220}$	1

QUESTION: What is the conditional pmf of Y_1, given $Y_2 = 1$?

SOLUTION. Straightforward calculations show that

$$\begin{aligned}
p_{Y_1|Y_2}(y_1 = 0|y_2 = 1) &= \frac{p_{Y_1,Y_2}(y_1 = 0, y_2 = 1)}{p_{Y_2}(y_2 = 1)} = \frac{40/220}{112/220} = 40/112 \\
p_{Y_1|Y_2}(y_1 = 1|y_2 = 1) &= \frac{p_{Y_1,Y_2}(y_1 = 1, y_2 = 1)}{p_{Y_2}(y_2 = 1)} = \frac{60/220}{112/220} = 60/112 \\
p_{Y_1|Y_2}(y_1 = 2|y_2 = 1) &= \frac{p_{Y_1,Y_2}(y_1 = 2, y_2 = 1)}{p_{Y_2}(y_2 = 1)} = \frac{12/220}{112/220} = 12/112.
\end{aligned}$$

Thus, the conditional pmf of Y_1, given $Y_2 = 1$, is given by

y_1	0	1	2
$p_{Y_1\|Y_2}(y_1\|y_2 = 1)$	40/112	60/112	12/112

This conditional pmf tells us how Y_1 is distributed if we are given that $Y_2 = 1$.

EXERCISE. Find the conditional pmf of Y_2, given $Y_1 = 0$. □

THE CONTINUOUS CASE: When (Y_1, Y_2) is a continuous random vector, we have to be careful how we define conditional distributions since the quantity

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_2}(y_2)}$$

has a zero denominator. As it turns out, this expression is the correct formula for the continuous case; however, we have to motivate its construction in a slightly different way.

ALTERNATE MOTIVATION: Suppose that (Y_1, Y_2) is a continuous random vector. For dy_1 and dy_2 small,

$$\begin{aligned}
f_{Y_1|Y_2}(y_1|y_2)dy_1 &= \frac{f_{Y_1,Y_2}(y_1, y_2)dy_1dy_2}{f_{Y_2}(y_2)dy_2} \\
&\approx \frac{P(y_1 \leq Y_1 \leq y_1 + dy_1, y_2 \leq Y_2 \leq y_2 + dy_2)}{P(y_2 \leq Y_2 \leq y_2 + dy_2)} \\
&= P(y_1 \leq Y_1 \leq y_1 + dy_1|y_2 \leq Y_2 \leq y_2 + dy_2).
\end{aligned}$$

Thus, we can think of $f_{Y_1|Y_2}(y_1|y_2)$ in this way; i.e., for "small" values of dy_1 and dy_2, $f_{Y_1|Y_2}(y_1|y_2)$ represents the conditional probability that Y_1 is between y_1 and $y_1 + dy_1$, given that Y_2 is between y_2 and $y_2 + dy_2$.

TERMINOLOGY: Suppose that (Y_1, Y_2) is a continuous random vector with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$. We define the **conditional probability density function (pdf)** of Y_1, given $Y_2 = y_2$, as

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_2}(y_2)}.$$

Similarly, the conditional probability density function of Y_2, given $Y_1 = y_1$, is

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_1}(y_1)}.$$

Example 4.5. Consider the bivariate pdf in Example 4.3:

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 6y_1, & 0 < y_1 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Recall that this probabilistic model summarized the random vector (Y_1, Y_2), where Y_1, the proportion of a population with Trait 1, is always less than Y_2, the proportion of a population with trait 2. Derive the conditional distributions $f_{Y_1|Y_2}(y_1|y_2)$ and $f_{Y_2|Y_1}(y_2|y_1)$.

Solution. In Example 4.3, we derived the marginal pdfs to be

$$f_{Y_1}(y_1) = \begin{cases} 6y_1(1 - y_1), & 0 < y_1 < 1 \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_{Y_2}(y_2) = \begin{cases} 3y_2^2, & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

First, we derive $f_{Y_1|Y_2}(y_1|y_2)$, so fix $Y_2 = y_2$. Remember, once we condition on $Y_2 = y_2$ (i.e., once we fix $Y_2 = y_2$), we then regard y_2 as simply some constant. **This is an important point to understand.** Then, for values of $0 < y_1 < y_2$, it follows that

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_2}(y_2)} = \frac{6y_1}{3y_2^2} = \frac{2y_1}{y_2^2},$$

and, thus, this is the value of $f_{Y_1|Y_2}(y_1|y_2)$ when $0 < y_1 < y_2$. Of course, for values of $y_1 \notin (0, y_2)$, the conditional density $f_{Y_1|Y_2}(y_1|y_2) = 0$. Summarizing, the conditional pdf of Y_1, given $Y_2 = y_2$, is given by

$$f_{Y_1|Y_2}(y_1|y_2) = \begin{cases} 2y_1/y_2^2, & 0 < y_1 < y_2 \\ 0, & \text{otherwise.} \end{cases}$$

Now, to derive the conditional pdf of Y_2 given Y_1, we fix $Y_1 = y_1$; then, for all values of $y_1 < y_2 < 1$, we have

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_1}(y_1)} = \frac{6y_1}{6y_1(1-y_1)} = \frac{1}{1-y_1}.$$

This is the value of $f_{Y_2|Y_1}(y_2|y_1)$ when $y_1 < y_2 < 1$. When $y_2 \notin (y_1, 1)$, the conditional pdf is $f_{Y_2|Y_1}(y_2|y_1) = 0$. Remember, once we condition on $Y_1 = y_1$, then we regard y_1 simply as some constant. Thus, the conditional pdf of Y_2, given $Y_1 = y_1$, is given by

$$f_{Y_2|Y_1}(y_2|y_1) = \begin{cases} \frac{1}{1-y_1}, & y_1 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

That is, conditional on $Y_1 = y_1$, $Y_2 \sim \mathcal{U}(y_1, 1)$. □

RESULT: The use of conditional densities allows us to define conditional probabilities of events associated with one random variable when we know the value of another random variable. If Y_1 and Y_2 are jointly **discrete**, then for any set $B \subset \mathcal{R}$,

$$P(Y_1 \in B|Y_2 = y_2) = \sum_B p_{Y_1|Y_2}(y_1|y_2).$$

If Y_1 and Y_2 are jointly **continuous**, then for any set $B \subset \mathcal{R}$,

$$P(Y_1 \in B|Y_2 = y_2) = \int_B f_{Y_1|Y_2}(y_1|y_2)dy_1.$$

Example 4.6. A small health-food store stocks two different brands of grain. Let Y_1 denote the amount of brand 1 in stock and let Y_2 denote the amount of brand 2 in stock (both Y_1 and Y_2 are measured in 100s of lbs). The joint distribution of Y_1 and Y_2 is given by

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} 24y_1y_2, & y_1 > 0,\ y_2 > 0,\ 0 < y_1 + y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the conditional pdf $f_{Y_1|Y_2}(y_1|y_2)$.

(b) Compute $P(Y_1 > 0.5|Y_2 = 0.3)$.

(c) Find $P(Y_1 > 0.5)$.

SOLUTIONS: (a) To find the conditional pdf $f_{Y_1|Y_2}(y_1|y_2)$, we first need to find the marginal pdf of Y_2. The marginal pdf of Y_2, for $0 < y_2 < 1$, is

$$f_{Y_2}(y_2) = \int_{y_1=0}^{1-y_2} 24y_1y_2 \, dy_1 = 24y_2 \left(\left. \frac{y_1^2}{2} \right|_0^{1-y_2} \right) = 12y_2(1-y_2)^2,$$

and 0, otherwise. Of course, we recognize this as a beta$(2,3)$ pdf; i.e., $Y_2 \sim$ beta$(2,3)$. The conditional pdf of Y_1, given $Y_2 = y_2$, is

$$\begin{aligned} f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_2}(y_2)} &= \frac{24y_1y_2}{12y_2(1-y_2)^2} \\ &= \frac{2y_1}{(1-y_2)^2}, \end{aligned}$$

for $0 < y_1 < 1 - y_2$, and 0, otherwise. Summarizing,

$$f_{Y_1|Y_2}(y_1|y_2) = \begin{cases} \frac{2y_1}{(1-y_2)^2}, & 0 < y_1 < 1 - y_2 \\ 0, & \text{otherwise.} \end{cases}$$

(b) To compute $P(Y_1 > 0.5|Y_2 = 0.3)$, we work with the conditional pdf $f_{Y_1|Y_2}(y_1|y_2)$, which for $y_2 = 0.3$, is given by

$$f_{Y_1|Y_2}(y_1|y_2) = \begin{cases} \left(\frac{200}{49}\right) y_1, & 0 < y_1 < 0.7 \\ 0, & \text{otherwise.} \end{cases}$$

Thus,

$$\begin{aligned} P(Y_1 > 0.5|Y_2 = 0.3) &= \int_{0.5}^{0.7} \left(\frac{200}{49}\right) y_1 dy_1 \\ &\approx 0.489. \end{aligned}$$

(c) To compute $P(Y_1 > 0.5)$, we can either use the marginal pdf $f_{Y_1}(y_1)$ or the joint pdf $f_{Y_1,Y_2}(y_1,y_2)$. Marginally, it turns out that $Y_1 \sim$ beta$(2,3)$ as well (verify!). Thus,

$$P(Y_1 > 0.5) = \int_{0.5}^{1} 12y_1(1-y_1)^2 dy_1 \approx 0.313.$$

REMARK: Notice how $P(Y_1 > 0.5|Y_2 = 0.3) \neq P(Y_1 > 0.5)$; that is, knowledge of the value of Y_2 has affected the way that we assign probability to events involving Y_1. Of course, one might expect this because of the support in the joint pdf $f_{Y_1,Y_2}(y_1,y_2)$. □

4.6 Independent random variables

TERMINOLOGY: Suppose that (Y_1, Y_2) is a random vector (discrete or continuous) with joint cdf $F_{Y_1,Y_2}(y_1, y_2)$, and denote the marginal cdfs of Y_1 and Y_2 by $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$, respectively. We say that the random variables Y_1 and Y_2 are **independent** if and only if

$$F_{Y_1,Y_2}(y_1, y_2) = F_{Y_1}(y_1)F_{Y_2}(y_2)$$

for all values of y_1 and y_2. Otherwise, we say that Y_1 and Y_2 are **dependent**.

RESULT: Suppose that (Y_1, Y_2) is a random vector (discrete or continuous) with joint pdf (pmf) $f_{Y_1,Y_2}(y_1, y_2)$, and denote the marginal pdfs (pmfs) of Y_1 and Y_2 by $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$, respectively. Then, Y_1 and Y_2 are independent if and only if

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1)f_{Y_2}(y_2)$$

for all values of y_1 and y_2. Otherwise, Y_1 and Y_2 are dependent.

Example 4.7. Suppose that the pmf for the discrete random vector (Y_1, Y_2) is given by

$$p_{Y_1,Y_2}(y_1, y_2) = \begin{cases} \frac{1}{18}(y_1 + 2y_2), & y_1 = 1, 2,\ y_2 = 1, 2 \\ 0, & \text{otherwise.} \end{cases}$$

The marginal distribution of Y_1, for values of $y_1 = 1, 2$, is given by

$$p_{Y_1}(y_1) = \sum_{y_2=1}^{2} p_{Y_1,Y_2}(y_1, y_2) = \sum_{y_2=1}^{2} \frac{1}{18}(y_1 + 2y_2) = \frac{1}{18}(2y_1 + 6),$$

and $p_{Y_1}(y_1) = 0$, otherwise. Similarly, the marginal distribution of Y_2, for values of $y_2 = 1, 2$, is given by

$$p_{Y_2}(y_2) = \sum_{y_1=1}^{2} p_{Y_1,Y_2}(y_1, y_2) = \sum_{y_1=1}^{2} \frac{1}{18}(y_1 + 2y_2) = \frac{1}{18}(3 + 4y_2),$$

and $p_{Y_2}(y_2) = 0$, otherwise. Note that, for example,

$$\frac{3}{18} = p_{Y_1,Y_2}(1, 1) \neq p_{Y_1}(1)p_{Y_2}(1) = \frac{8}{18} \times \frac{7}{18} = \frac{14}{81};$$

thus, the random variables Y_1 and Y_2 are dependent. □

Example 4.8. Let Y_1 and Y_2 denote the proportions of time (out of one workday) during which employees I and II, respectively, perform their assigned tasks. Suppose that the random vector (Y_1, Y_2) has joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} y_1 + y_2, & 0 < y_1 < 1,\ 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

It is straightforward to show (verify!) that

$$f_{Y_1}(y_1) = \begin{cases} y_1 + \frac{1}{2}, & 0 < y_1 < 1 \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_{Y_2}(y_2) = \begin{cases} y_2 + \frac{1}{2}, & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Thus, since $f_{Y_1,Y_2}(y_1, y_2) = y_1 + y_2 \neq (y_1 + \frac{1}{2})(y_2 + \frac{1}{2}) = f_{Y_1}(y_1) f_{Y_2}(y_2)$, for $0 < y_1 < 1$ and $0 < y_2 < 1$, Y_1 and Y_2 are dependent. □

Example 4.9. Suppose that Y_1 and Y_2 represent the death times (in hours) for rats treated with a certain toxin. Marginally, each death time follows an exponential distribution with mean θ, and Y_1 and Y_2 are independent.

(a) Write out the joint pdf of (Y_1, Y_2).

(b) Compute $P(Y_1 \leq 1, Y_2 \leq 1)$.

SOLUTIONS: (a) Because Y_1 and Y_2 are independent, the joint pdf of (Y_1, Y_2), for $y_1 > 0$ and $y_2 > 0$, is given by

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2}(y_2) = \frac{1}{\theta} e^{-y_1/\theta} \times \frac{1}{\theta} e^{-y_2/\theta} = \frac{1}{\theta^2} e^{-(y_1+y_2)/\theta},$$

and $f_{Y_1,Y_2}(y_1, y_2) = 0$ otherwise.

(b) Because Y_1 and Y_2 are independent,

$$\begin{aligned} P(Y_1 \leq 1, Y_2 \leq 1) = F_{Y_1,Y_2}(1, 1) &= F_{Y_1}(1) F_{Y_2}(1) \\ &= (1 - e^{-1/\theta})(1 - e^{-1/\theta}) \\ &= (1 - e^{-1/\theta})^2. \ \square \end{aligned}$$

A CONVENIENT RESULT: Let (Y_1, Y_2) be a random vector (discrete or continuous) with pdf (pmf) $f_{Y_1,Y_2}(y_1, y_2)$, If the support set R_{Y_1,Y_2} does not constrain y_1 by y_2 (or y_2 by y_1), and additionally, we can factor the joint pdf (pmf) $f_{Y_1,Y_2}(y_1, y_2)$ into two nonnegative expressions

$$f_{Y_1,Y_2}(y_1, y_2) = g(y_1)h(y_2),$$

then Y_1 and Y_2 are independent. Note that $g(y_1)$ and $h(y_2)$ are simply functions; **they need not be pdfs (pmfs)**, although they sometimes are. The only requirement is that $g(y_1)$ is a function of y_1 only, $h(y_2)$ is a function of y_2 only, and that both are nonnegative. *If the support involves a constraint, the random variables are automatically dependent.*

Example 4.10. In Example 4.6, Y_1 denoted the amount of brand 1 grain in stock and Y_2 denoted the amount of brand 2 grain in stock. Recall that the joint pdf of (Y_1, Y_2) was given by

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 24y_1y_2, & y_1 > 0,\ y_2 > 0,\ 0 < y_1 + y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Here, the support is $R_{Y_1,Y_2} = \{(y_1, y_2) : y_1 > 0,\ y_2 > 0,\ 0 < y_1 + y_2 < 1\}$. Since knowledge of y_1 (y_2) affects the value of y_2 (y_1), the support involves a constraint, and Y_1 and Y_2 are dependent. □

Example 4.11. Suppose that the random vector (X, Y) has joint pdf

$$f_{X,Y}(x, y) = \begin{cases} [\Gamma(\alpha)\Gamma(\beta)]^{-1}\lambda e^{-\lambda x}(\lambda x)^{\alpha+\beta-1}y^{\alpha-1}(1-y)^{\beta-1}, & x > 0, 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

for $\lambda > 0$, $\alpha > 0$, and $\beta > 0$. Since the support $R_{X,Y} = \{(x, y) : x > 0,\ 0 < y < 1\}$ does not involve a constraint, it follows immediately that X and Y are independent, since we can write

$$f_{X,Y}(x, y) = \underbrace{\lambda e^{-\lambda x}(\lambda x)^{\alpha+\beta-1}}_{g(x)} \times \underbrace{\frac{y^{\alpha-1}(1-y)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}}_{h(y)}.$$

Note that we are not saying that $g(x)$ and $h(y)$ are marginal distributions of X and Y, respectively (in fact, they are **not** the marginal distributions). □

EXTENSION: We generalize the notion of **independence** to n-variate random vectors. We use the conventional notation $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ and $\boldsymbol{y} = (y_1, y_2, ..., y_n)$. Also, we will denote the joint cdf of $\boldsymbol{Y}$ by $F_{\boldsymbol{Y}}(\boldsymbol{y})$ and the joint pdf (pmf) of $\boldsymbol{Y}$ by $f_{\boldsymbol{Y}}(\boldsymbol{y})$.

TERMINOLOGY: Suppose that the random vector $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ has joint cdf $F_{\boldsymbol{Y}}(\boldsymbol{y})$, and suppose that the random variable Y_i has cdf $F_{Y_i}(y_i)$, for $i = 1, 2, ..., n$. Then, $Y_1, Y_2, ..., Y_n$ are **independent** random variables if and only if

$$F_{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{i=1}^{n} F_{Y_i}(y_i);$$

that is, the joint cdf can be factored into the product of the marginal cdfs. Alternatively, $Y_1, Y_2, ..., Y_n$ are **independent** random variables if and only if

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i);$$

that is, the joint pdf (pmf) can be factored into the product of the marginals.

Example 4.12. In a small clinical trial, $n = 20$ patients are treated with a new drug. Suppose that the response from each patient is a measurement $Y \sim \mathcal{N}(\mu, \sigma^2)$. Denoting the 20 responses by $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_{20})$, then, assuming independence, the joint distribution of the 20 responses is, for $\boldsymbol{y} \in \mathcal{R}^{20}$,

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{i=1}^{20} \underbrace{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2}}_{f_{Y_i}(y_i)} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{20} e^{-\frac{1}{2}\sum_{i=1}^{20}\left(\frac{y_i-\mu}{\sigma}\right)^2}.$$

What is the probability that every patient's response is less than $\mu + 2\sigma$?

Solution: The probability that Y_1 is less than $\mu + 2\sigma$ is given by

$$P(Y_1 < \mu + 2\sigma) = P(Z < 2) = \Phi(2) = 0.9772,$$

where $Z \sim \mathcal{N}(0, 1)$ and $\Phi(\cdot)$ denotes the standard normal cdf. Because the patients' responses are independent random variables,

$$\begin{aligned} P(Y_1 < \mu + 2\sigma, Y_2 < \mu + 2\sigma, ..., Y_{20} < \mu + 2\sigma) &= \prod_{i=1}^{20} P(Y_i < \mu + 2\sigma) \\ &= [\Phi(2)]^{20} \approx 0.630. \quad \square \end{aligned}$$

4.7 Expectations of functions of random variables

RESULT: Suppose that $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ has joint pdf $f_{\boldsymbol{Y}}(\boldsymbol{y})$, or joint pmf $p_{\boldsymbol{Y}}(\boldsymbol{y})$, and suppose that $g(\boldsymbol{Y}) = g(Y_1, Y_2, ..., Y_n)$ is any real vector valued function of $Y_1, Y_2, ..., Y_n$; i.e., $g : \mathcal{R}^n \rightarrow \mathcal{R}$. Then,

- if $\boldsymbol{Y}$ is discrete,

$$E[g(\boldsymbol{Y})] = \sum_{\text{all } y_1} \sum_{\text{all } y_2} \cdots \sum_{\text{all } y_n} g(\boldsymbol{y}) p_{\boldsymbol{Y}}(\boldsymbol{y}),$$

- and if $\boldsymbol{Y}$ is continuous,

$$E[g(\boldsymbol{Y})] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\boldsymbol{y}) f_{\boldsymbol{Y}}(\boldsymbol{y}) d\boldsymbol{y}.$$

If these quantities are not finite, then we say that $E[g(\boldsymbol{Y})]$ does not exist.

Example 4.13. In Example 4.6, Y_1 denotes the amount of grain 1 in stock and Y_2 denotes the amount of grain 2 in stock. The joint distribution of Y_1 and Y_2 was given by

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 24y_1y_2, & y_1 > 0,\ y_2 > 0,\ 0 < y_1 + y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

What is the expected **total** amount of grain $(Y_1 + Y_2)$ in stock?

Solution: Let the function $g : \mathcal{R}^2 \rightarrow \mathcal{R}$ be defined by $g(y_1, y_2) = y_1 + y_2$. We would like to compute $E[g(Y_1, Y_2)] = E(Y_1 + Y_2)$. From the last result, we know that

$$\begin{aligned} E(Y_1 + Y_2) &= \int_{y_1=0}^{1} \int_{y_2=0}^{1-y_1} (y_1 + y_2) 24y_1y_2 \ dy_2 dy_1 \\ &= \int_{y_1=0}^{1} \left[\left(24y_1^2 \frac{y_2^2}{2} \Big|_0^{1-y_1} \right) + \left(24y_1 \frac{y_2^3}{3} \Big|_0^{1-y_1} \right) \right] dy_1 \\ &= \int_{y_1=0}^{1} 12y_1^2(1-y_1)^2 dy_1 + \int_{y_1=0}^{1} 8y_1(1-y_1)^3 dy_1 \\ &= 12 \left[\frac{\Gamma(3)\Gamma(3)}{\Gamma(6)} \right] + 8 \left[\frac{\Gamma(2)\Gamma(4)}{\Gamma(6)} \right] = 4/5. \end{aligned}$$

The expected amount of grain in stock is 80 lbs. Recall that, marginally, $Y_1 \sim \text{beta}(2, 3)$ and $Y_2 \sim \text{beta}(2, 3)$ so that $E(Y_1) = E(Y_2) = \frac{2}{5}$ and $E(Y_1 + Y_2) = \frac{2}{5} + \frac{2}{5} = \frac{4}{5}$. □

Example 4.14. A process for producing an industrial chemical yields a product containing two types of impurities (Type I and Type II). From a specified sample from this process, let Y_1 denote the proportion of impurities in the sample (of both types) and let Y_2 denote the proportion of Type I impurities among all impurities found. Suppose that the joint pdf of the random vector (Y_1, Y_2) is given by

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 2(1-y_1), & 0 < y_1 < 1,\ 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Find the expected value of the proportion of Type I impurities in the sample.

Solution: Because Y_1 is the proportion of impurities in the sample and Y_2 is the proportion of Type I impurities among the sample impurities, it follows that Y_1Y_2 is the proportion of Type I impurities in the sample taken. Let the function $g : \mathcal{R}^2 \to \mathcal{R}$ be defined by $g(y_1, y_2) = y_1y_2$. We would like to compute $E[g(Y_1, Y_2)] = E(Y_1Y_2)$. This is given by

$$E(Y_1Y_2) = \int_0^1 \int_0^1 y_1y_2 2(1-y_1)dy_1dy_2 = \frac{1}{6}. \ \square$$

PROPERTIES OF EXPECTATIONS: Let $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ be a discrete or continuous random vector with pdf (pmf) $f_{\boldsymbol{Y}}(\boldsymbol{y})$ and support $R \subset \mathcal{R}^n$, suppose that $g, g_1, g_2, ..., g_k$ are real vector valued functions from $\mathcal{R}^n \to \mathcal{R}$, and let c be any real constant. Then,

(a) $E(c) = c$

(b) $E[cg(\boldsymbol{Y})] = cE[g(\boldsymbol{Y})]$

(c) $E[\sum_{j=1}^k g_j(\boldsymbol{Y})] = \sum_{j=1}^k E[g_j(\boldsymbol{Y})]$.

RESULT: Suppose that Y_1 and Y_2 are **independent** random variables, and consider the functions $g(Y_1)$ and $h(Y_2)$, where $g(Y_1)$ is a function of Y_1 only, and $h(Y_2)$ is a function of Y_2 only. Then,

$$E[g(Y_1)h(Y_2)] = E[g(Y_1)]E[h(Y_2)],$$

provided that all expectations exist.

Proof. Without loss, we will assume that (Y_1, Y_2) is a continuous random vector (the discrete case is analogous). Suppose that (Y_1, Y_2) has joint pdf $f_{Y_1,Y_2}(y_1, y_2)$ with support $R \subset \mathcal{R}^2$. Note that

$$\begin{aligned} E[g(Y_1)h(Y_2)] &= \int_{\mathcal{R}^2} g(y_1)h(y_2)f_{Y_1,Y_2}(y_1,y_2)dy_2dy_1 \\ &= \int_{\mathcal{R}}\int_{\mathcal{R}} g(y_1)h(y_2)f_{Y_1}(y_1)f_{Y_2}(y_2)dy_2dy_1 \\ &= \int_{\mathcal{R}} g(y_1)f_{Y_1}(y_1)dy_1 \left[\int_{\mathcal{R}} h(y_2)f_{Y_2}(y_2)dy_2\right] \\ &= E[h(Y_2)]\int_{\mathcal{R}} g(y_1)f_{Y_1}(y_1)dy_1 \\ &= E[h(Y_2)]E[g(Y_1)]. \ \square \end{aligned}$$

Example 4.15. A point $(Y_1, Y_2) \in \mathcal{R}^2$ is selected at random, where $Y_1 \sim \mathcal{N}(\mu_1, \sigma^2)$, $Y_2 \sim \mathcal{N}(\mu_2, \sigma^2)$, and Y_1 and Y_2 are independent. Define the random variables

$$\begin{aligned} T &= Y_1 + Y_2 \\ U &= Y_1Y_2 \\ Z &= Y_1^2 + Y_2^2. \end{aligned}$$

Find $E(T)$, $E(U)$, and $E(Z)$.

SOLUTIONS. (a) Because $E(\cdot)$ is linear, we know

$$E(T) = E(Y_1 + Y_2) = E(Y_1) + E(Y_2) = \mu_1 + \mu_2.$$

Because Y_1 and Y_2 are independent, we know that

$$E(U) = E(Y_1Y_2) = E(Y_1)E(Y_2) = \mu_1\mu_2.$$

To compute $E(Z)$, first note that

$$E(Y_1^2) = V(Y_1) + [E(Y_1)]^2 = \sigma^2 + \mu_1^2$$

and

$$E(Y_2^2) = V(Y_2) + [E(Y_2)]^2 = \sigma^2 + \mu_2^2$$

so that

$$\begin{aligned} E(Z) = E(Y_1^2 + Y_2^2) = E(Y_1^2) + E(Y_2^2) &= (\sigma^2 + \mu_1^2) + (\sigma^2 + \mu_2^2) \\ &= 2\sigma^2 + \mu_1^2 + \mu_2^2. \ \square \end{aligned}$$

EXERCISE: Compute $E(TU)$, $E(TZ)$, and $E(UZ)$.

4.8 Covariance and correlation

4.8.1 Covariance

TERMINOLOGY: Suppose that Y_1 and Y_2 are random variables with means μ_{Y_1} and μ_{Y_2}, respectively. The **covariance** between Y_1 and Y_2 is given by

$$\text{Cov}(Y_1, Y_2) = E[(Y_1 - \mu_{Y_1})(Y_2 - \mu_{Y_2})].$$

The covariance gives us information about how Y_1 and Y_2 are **linearly** related.

THE COVARIANCE COMPUTING FORMULA: It is easy to show that

$$\text{Cov}(Y_1, Y_2) \equiv E[(Y_1 - \mu_{Y_1})(Y_2 - \mu_{Y_2})] = E(Y_1 Y_2) - \mu_{Y_1}\mu_{Y_2}.$$

This latter expression is sometimes easier to work with and is called the **covariance computing formula**.

Example 4.16. Gasoline is stocked in a tank once at the beginning of each week and then sold to customers. Let Y_1 denote the proportion of the capacity of the tank that is available after it is stocked. Let Y_2 denote the proportion of the capacity of the bulk tank that is sold during the week. Suppose that the random vector (Y_1, Y_2) has joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 3y_1, & 0 < y_2 < y_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

To compute the covariance, first note that $Y_1 \sim \text{beta}(3, 1)$ and $Y_2 \sim f_{Y_2}(y_2)$, where

$$f_{Y_2}(y_2) = \begin{cases} \frac{3}{2}(1 - y_2^2), & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Thus, $E(Y_1) = 3/(3 + 1) = 0.75$ and

$$E(Y_2) = \int_0^1 y_2 \times \frac{3}{2}(1 - y_2^2) dy = 0.375.$$

Also,

$$E(Y_1 Y_2) = \int_{y_1=0}^1 \int_{y_2=0}^{y_1} y_1 y_2 \times 3y_1 dy_2 dy_1 = 0.30.$$

Thus, the covariance is

$$\begin{aligned}\text{Cov}(Y_1, Y_2) &= E(Y_1Y_2) - \mu_{Y_1}\mu_{Y_2} \\ &= 0.30 - (0.75)(0.375) = 0.01875. \ \square\end{aligned}$$

NOTES ON THE COVARIANCE:

- If $\text{Cov}(Y_1, Y_2) > 0$, then Y_1 and Y_2 are **positively** linearly related.
- If $\text{Cov}(Y_1, Y_2) < 0$, then Y_1 and Y_2 are **negatively** linearly related.
- If $\text{Cov}(Y_1, Y_2) = 0$, then Y_1 and Y_2 are not linearly related. This does **not** necessarily mean that Y_1 and Y_2 are independent!

RESULT: If Y_1 and Y_2 are independent, then $\text{Cov}(Y_1, Y_2) = 0$.
Proof. Using the covariance computing formula, we have

$$\begin{aligned}\text{Cov}(Y_1, Y_2) &= E(Y_1Y_2) - \mu_{Y_1}\mu_{Y_2} \\ &= E(Y_1)E(Y_2) - \mu_{Y_1}\mu_{Y_2} = 0. \ \square\end{aligned}$$

MAIN POINT: If two random variables are independent, then they have zero covariance; however, zero covariance does not necessarily imply independence.

Example 4.17. *An example of two dependent variables with zero covariance.* Suppose that $Y_1 \sim \mathcal{U}(-1, 1)$, and let $Y_2 = Y_1^2$. It is straightforward to show that $E(Y_1) = 0$, $E(Y_1Y_2) = E(Y_1^3) = 0$, and $E(Y_2) = E(Y_1^2) = V(Y_1) = 1/3$. Thus,

$$\text{Cov}(Y_1, Y_2) = E(Y_1Y_2) - \mu_{Y_1}\mu_{Y_2} = 0 - 0(1/3) = 0.$$

However, not only are Y_1 and Y_2 related, they are **perfectly** related! But, the relationship is not linear (it is quadratic). The covariance only assesses linear relationships. $\square$

IMPORTANT RESULT: Suppose that Y_1 and Y_2 are random variables. Then,

$$\begin{aligned}V(Y_1 + Y_2) &= V(Y_1) + V(Y_2) + 2\text{Cov}(Y_1, Y_2) \\ V(Y_1 - Y_2) &= V(Y_1) + V(Y_2) - 2\text{Cov}(Y_1, Y_2).\end{aligned}$$

Proof. Let $Z = Y_1 + Y_2$. Using the definition of variance, we have

$$\begin{aligned}
V(Z) &= E[(Z-\mu_Z)^2] \\
&= E\{[(Y_1+Y_2) - E(Y_1+Y_2)]^2\} \\
&= E[(Y_1+Y_2-\mu_{Y_1}-\mu_{Y_2})^2] \\
&= E\{[(Y_1-\mu_{Y_1}) + (Y_2-\mu_{Y_2})]^2\} \\
&= E[(Y_1-\mu_{Y_1})^2 + (Y_2-\mu_{Y_2})^2 + 2\underbrace{(Y_1-\mu_{Y_1})(Y_2-\mu_{Y_2})}_{\text{cross product}}] \\
&= E[(Y_1-\mu_{Y_1})^2] + E[(Y_2-\mu_{Y_2})^2] + 2E[(Y_1-\mu_{Y_1})(Y_2-\mu_{Y_2})] \\
&= V(Y_1) + V(Y_2) + 2\text{Cov}(Y_1,Y_2).
\end{aligned}$$

That $V(Y_1 - Y_2) = V(Y_1) + V(Y_2) - 2\text{Cov}(Y_1, Y_2)$ is shown similarly. □

Example 4.18. A small health-food store stocks two different brands of grain. Let Y_1 denote the amount of brand 1 in stock and let Y_2 denote the amount of brand 2 in stock (both Y_1 and Y_2 are measured in 100s of lbs). In Example 4.6, we saw that the joint distribution of Y_1 and Y_2 was given by

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} 24y_1y_2, & y_1>0,\ y_2>0,\ 0<y_1+y_2<1 \\ 0, & \text{otherwise.}\end{cases}$$

What is the variance for the **total** amount of grain in stock? That is, what is $V(Y_1+Y_2)$?

SOLUTION: Using the last result, we know that

$$V(Y_1+Y_2) = V(Y_1) + V(Y_2) + 2\text{Cov}(Y_1,Y_2).$$

Marginally, Y_1 and Y_2 both have beta$(2, 3)$ distributions (see Example 4.6). Thus,

$$E(Y_1) = E(Y_2) = \frac{2}{2+3} = \frac{2}{5}.$$

and

$$V(Y_1) = V(Y_2) = \frac{2(3)}{(2+3+1)(2+3)^2} = \frac{1}{25}.$$

Recall that $\text{Cov}(Y_1,Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2)$, so we need to first compute $E(Y_1Y_2)$:

$$E(Y_1Y_2) = \int_{y_1=0}^{1}\int_{y_2=0}^{1-y_1} y_1y_2 \times 24y_1y_2\ dy_2dy_1 = \frac{2}{15}.$$

Thus,

$$\text{Cov}(Y_1, Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2) = \frac{2}{15} - \left(\frac{2}{5}\right)\left(\frac{2}{5}\right) \approx -0.027.$$

Finally, the variance of $Y_1 + Y_2$ is given by

$$V(Y_1 + Y_2) = \frac{1}{25} + \frac{1}{25} + 2(-0.027) \approx 0.027. \quad \square$$

RESULT: Suppose that Y_1 and Y_2 are **independent** random variables. Then,

$$V(Y_1 \pm Y_2) = V(Y_1) + V(Y_2).$$

Proof. In general, $V(Y_1 \pm Y_2) = V(Y_1) + V(Y_2) \pm 2\text{Cov}(Y_1, Y_2)$. Since Y_1 and Y_2 are independent, $\text{Cov}(Y_1, Y_2) = 0$. Thus, the result follows immediately. □

LEMMA: Suppose that Y_1 and Y_2 are random variables with means μ_{Y_1} and μ_{Y_2}, respectively. Then,

(a) $\text{Cov}(Y_1, Y_2) = \text{Cov}(Y_2, Y_1)$

(b) $\text{Cov}(Y_1, Y_1) = V(Y_1)$.

(c) $\text{Cov}(a + bY_1, c + dY_2) = bd\text{Cov}(Y_1, Y_2)$, for constants a, b, c, and d.

Proof. Exercise. □

4.8.2 Correlation

GENERAL PROBLEM: Suppose that X and Y are random variables and that we want to predict Y as a **linear function** of X. That is, we want to consider functions of the form $Y = \beta_0 + \beta_1 X$, for constants β_0 and β_1. In this situation, the "error in prediction" is given by

$$Y - (\beta_0 + \beta_1 X).$$

This error can be positive or negative, so in developing a "goodness measure" of prediction error, we want one that maintains the magnitude of error but ignores the sign. Thus, consider the **mean squared error of prediction** given by

$$Q(\beta_0, \beta_1) \equiv E\{[Y - (\beta_0 + \beta_1 X)]^2\}.$$

A two-variable calculus argument shows that the mean squared error of prediction $Q(\beta_0, \beta_1)$ is minimized when

$$\beta_1 = \frac{\text{Cov}(X, Y)}{V(X)}$$

and

$$\beta_0 = E(Y) - \left[\frac{\text{Cov}(X, Y)}{V(X)}\right] E(X).$$

However, note that the value of β_1, algebraically, is equal to

$$\begin{aligned}\beta_1 &= \frac{\text{Cov}(X, Y)}{V(X)} \\ &= \left[\frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}\right] \frac{\sigma_Y}{\sigma_X} \\ &= \rho_{X,Y}\left(\frac{\sigma_Y}{\sigma_X}\right),\end{aligned}$$

where

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

The quantity $\rho_{X,Y}$ is called the **correlation coefficient** between X and Y.

SUMMARY: The best **linear predictor** of Y, given X, is $Y = \beta_0 + \beta_1 X$, where

$$\begin{aligned}\beta_1 &= \rho_{X,Y}\left(\frac{\sigma_Y}{\sigma_X}\right) \\ \beta_0 &= E(Y) - \beta_1 E(X).\end{aligned}$$

NOTES ON THE CORRELATION COEFFICIENT:

(1) $-1 \leq \rho_{X,Y} \leq 1$ (this can be proven using the Cauchy-Schwartz Inequality, from calculus).

(2) If $\rho_{X,Y} = 1$, then $Y = \beta_0 + \beta_1 X$, where $\beta_1 > 0$. That is, X and Y are **perfectly positively linearly** related; i.e., the bivariate probability distribution of (X, Y) lies entirely on a straight line with positive slope.

(3) If $\rho_{X,Y} = -1$, then $Y = \beta_0 + \beta_1 X$, where $\beta_1 < 0$. That is, X and Y are **perfectly negatively linearly** related; i.e., the bivariate probability distribution of (X, Y) lies entirely on a straight line with negative slope.

(4) If $\rho_{X,Y} = 0$, then X and Y are not **linearly** related.

NOTE: If X and Y are independent random variables, then $\rho_{X,Y} = 0$. However, again, the implication does not go the other way; that is, if $\rho_{X,Y} = 0$, this does not necessarily mean that X and Y are independent.

NOTE: In assessing the strength of the linear relationship between X and Y, the correlation coefficient is often preferred over the covariance since $\rho_{X,Y}$ is measured on a bounded, unitless scale. On the other hand, $\text{Cov}(X, Y)$ can be any real number.

Example 4.19. In Example 4.16, we considered the bivariate model

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 3y_1, & 0 < y_2 < y_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

for Y_1, the proportion of the capacity of the tank after being stocked, and Y_2, the proportion of the capacity of the tank that is sold. What is ρ_{Y_1,Y_2}?

SOLUTION: In Example 4.16, we computed $\text{Cov}(Y_1, Y_2) = 0.01875$, so all we need is σ_{Y_1} and σ_{Y_2}. We also found that $Y_1 \sim \text{beta}(3, 1)$ and $Y_2 \sim f_{Y_2}(y_2)$, where

$$f_{Y_2}(y_2) = \begin{cases} \frac{3}{2}(1 - y_2^2), & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

The variance of Y_1 is

$$V(Y_1) = \frac{3(1)}{(3+1+1)(3+1)^2} = \frac{3}{80} \implies \sigma_{Y_1} = \sqrt{\frac{3}{80}} \approx 0.194.$$

Simple calculations using $f_{Y_2}(y_2)$ show that $E(Y_2^2) = 1/5$ and $E(Y_2) = 3/8$ so that

$$V(Y_2) = \frac{1}{5} - \left(\frac{3}{8}\right)^2 = 0.059 \implies \sigma_{Y_2} = \sqrt{0.059} \approx 0.244.$$

Thus,

$$\rho_{Y_1,Y_2} = \frac{\text{Cov}(Y_1, Y_2)}{\sigma_{Y_1}\sigma_{Y_2}} \approx \frac{0.01875}{0.194 \times 0.244} \approx 0.40. \ \square$$

4.9 Expectations and variances of linear functions of random variables

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ are random variables and that $a_1, a_2, ..., a_n$ are constants. The function

$$U = \sum_{i=1}^{n} a_i Y_i = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$$

is called a **linear combination** of the random variables $Y_1, Y_2, ..., Y_n$.

EXPECTED VALUE OF A LINEAR COMBINATION:

$$E(U) = E\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i E(Y_i)$$

VARIANCE OF A LINEAR COMBINATION:

$$\begin{aligned} V(U) = V\left(\sum_{i=1}^{n} a_i Y_i\right) &= \sum_{i=1}^{n} a_i^2 V(Y_i) + 2\sum_{i<j} a_i a_j \text{Cov}(Y_i, Y_j) \\ &= \sum_{i=1}^{n} a_i^2 V(Y_i) + \sum_{i \neq j} a_i a_j \text{Cov}(Y_i, Y_j) \end{aligned}$$

COVARIANCE BETWEEN TWO LINEAR COMBINATIONS: Suppose that

$$\begin{aligned} U_1 &= \sum_{i=1}^{n} a_i Y_i = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n \\ U_2 &= \sum_{j=1}^{m} b_j X_j = b_1 X_1 + b_2 X_2 + \cdots + b_m X_m. \end{aligned}$$

Then, it follows that

$$\text{Cov}(U_1, U_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \text{Cov}(Y_i, X_j).$$

BIVARIATE CASE: Interest will often focus on situations wherein we have a linear combination of $n = 2$ random variables. In this setting,

$$\begin{aligned} E(a_1 Y_1 + a_2 Y_2) &= a_1 E(Y_1) + a_2 E(Y_2) \\ V(a_1 Y_1 + a_2 Y_2) &= a_1^2 V(Y_1) + a_2^2 V(Y_2) + 2 a_1 a_2 \text{Cov}(Y_1, Y_2). \end{aligned}$$

Similarly, when $n = m = 2$,

$$\begin{aligned}\text{Cov}(a_1Y_1 + a_2Y_2, b_1X_1 + b_2X_2) &= a_1b_1\text{Cov}(Y_1, X_1) + a_1b_2\text{Cov}(Y_1, X_2)\\ &\quad + a_2b_1\text{Cov}(Y_2, X_1) + a_2b_2\text{Cov}(Y_2, X_2).\end{aligned}$$

Example 4.20. Achievement tests are usually seen in educational or employment settings. These tests attempt to measure how much you know about a certain topic in a particular area. Suppose that Y_1, Y_2, and Y_3 represent scores for a particular different parts of an exam. It is posited that $Y_1 \sim \mathcal{N}(12, 4)$, $Y_2 \sim \mathcal{N}(16, 9)$, $Y_3 \sim \mathcal{N}(20, 16)$, Y_1 and Y_2 are independent, $\text{Cov}(Y_1, Y_3) = 0.8$, and $\text{Cov}(Y_2, Y_3) = -6.7$. Two different summary measures are computed to assess a subject's performance:

$$U_1 = 0.5Y_1 - 2Y_2 + Y_3 \quad \text{and} \quad U_2 = 3Y_1 - 2Y_2 - Y_3.$$

(a) $E(U_1)$ and $V(U_1)$.

(b) Find $\text{Cov}(U_1, U_2)$.

Solutions: The **mean** of U_1 is

$$\begin{aligned}E(U_1) = E(0.5Y_1 - 2Y_2 + Y_3) &= 0.5E(Y_1) - 2E(Y_2) + E(Y_3)\\ &= 0.5(12) - 2(16) + 20 = -6.\end{aligned}$$

The **variance** of U_1 is

$$\begin{aligned}V(U_1) &= V(0.5Y_1 - 2Y_2 + Y_3)\\ &= (0.5)^2V(Y_1) + (-2)^2V(Y_2) + V(Y_3)\\ &\quad + 2(0.5)(-2)\text{Cov}(Y_1, Y_2) + 2(0.5)(1)\text{Cov}(Y_1, Y_3) + 2(-2)(1)\text{Cov}(Y_2, Y_3)\\ &= (0.25)(4) + 4(9) + 16 + 2(0.5)(-2)(0) + 2(0.5)(0.8) + 2(-2)(-6.7) = 80.6.\end{aligned}$$

The **covariance** between U_1 and U_2 is

$$\begin{aligned}\text{Cov}(U_1, U_2) &= \text{Cov}(0.5Y_1 - 2Y_2 + Y_3,\ 3Y_1 - 2Y_2 - Y_3)\\ &= (0.5)(3)\text{Cov}(Y_1, Y_1) + (0.5)(-2)\text{Cov}(Y_1, Y_2) + (0.5)(-1)\text{Cov}(Y_1, Y_3)\\ &\quad + (-2)(3)\text{Cov}(Y_2, Y_1) + (-2)(-2)\text{Cov}(Y_2, Y_2) + (-2)(-1)\text{Cov}(Y_2, Y_3)\\ &\quad + (1)(3)\text{Cov}(Y_3, Y_1) + (1)(-2)\text{Cov}(Y_3, Y_2) + (1)(-1)\text{Cov}(Y_3, Y_3)\\ &= 28. \ \square\end{aligned}$$

4.10 The multinomial model

RECALL: When we discussed the binomial model in Chapter 2, each Bernoulli trial resulted in either a "success" or a "failure;" that is, on each trial, there were only two outcomes possible (e.g., infected/not, germinated/not, defective/not, etc.).

TERMINOLOGY: A **multinomial experiment** is simply a generalization of a binomial experiment. In particular, consider an experiment where

- the experiment consists of n trials (n is fixed),
- the outcome for any trial belongs to **exactly** one of $k \geq 2$ classes,
- the probability that an outcome for a single trial falls into class i is given by p_i, for $i = 1, 2, ..., k$, where each p_i remains constant from trial to trial, and
- trials are independent.

DEFINITION: In a multinomial experiment, let Y_i denote the number of outcomes in class i, so that $Y_1 + Y_2 + \cdots + Y_k = n$, and denote $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_k)$. We call $\boldsymbol{Y}$ a **multinomial** random vector and write $\boldsymbol{Y} \sim \text{mult}(n, p_1, p_2, ..., p_k; \sum_i p_i = 1)$.

NOTE: When $k = 2$, the multinomial random vector reduces to our well-known binomial situation. When $k = 3$, $\boldsymbol{Y}$ would be called a **trinomial** random vector.

JOINT PMF: If $\boldsymbol{Y} \sim \text{mult}(n, p_1, p_2, ..., p_k; \sum_i p_i = 1)$, the pmf for $\boldsymbol{Y}$ is given by

$$p_{\boldsymbol{Y}}(\boldsymbol{y}) = \begin{cases} \frac{n!}{y_1!y_2!\cdots y_k!}p_1^{y_1}p_2^{y_2}\cdots p_k^{y_k}, & y_i = 0, 1, ..., n; \sum_i y_i = n \\ 0, & \text{otherwise.} \end{cases}$$

Example 4.21. In a manufacturing experiment, we observe $n = 10$ parts, each of which can be classified as non-defective, defective, or reworkable. Define

$$\begin{aligned} Y_1 &= \text{number of non-defective parts} \\ Y_2 &= \text{number of defective parts} \\ Y_3 &= \text{number of reworkable parts.} \end{aligned}$$

Assuming that each part (i.e., trial) is independent of other parts, a multinomial model applies and $\boldsymbol{Y} = (Y_1, Y_2, Y_3) \sim \text{mult}(10, p_1, p_2, p_3; \sum_i p_i = 1)$. Suppose that $p_1 = 0.90$, $p_2 = 0.03$, and $p_3 = 0.07$. What is the probability that a sample (of 10) contains 8 non-defective parts, 1 defective part, and 1 reworkable part?

SOLUTION: We want to compute $p_{Y_1,Y_2,Y_3}(8, 1, 1)$. This equals

$$p_{Y_1,Y_2,Y_3}(8, 1, 1) = \frac{10!}{8!1!1!}(0.90)^8(0.03)^1(0.07)^1 \approx 0.081. \quad \square$$

Example 4.22. At a number of clinic sites throughout Nebraska, chlamydia and gonorrhea testing is performed on individuals using urine or cervical-swab specimens. More than 30,000 of these tests are done annually by the Nebraska Public Health Laboratory! Suppose that on a given day, there are $n = 280$ subjects tested, and define

$$\begin{aligned} p_1 &= \text{proportion of subjects with neither chlamydia nor gonorrhea} \\ p_2 &= \text{proportion of subjects with chlamydia but not gonorrhea} \\ p_3 &= \text{proportion of subjects with gonorrhea but not chlamydia} \\ p_4 &= \text{proportion of subjects with both chlamydia and gonorrhea.} \end{aligned}$$

Define $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)$, where Y_i counts the number of subjects in category i. Assuming that subjects are independent, $\boldsymbol{Y} \sim \text{mult}(280, p_1, p_2, p_3, p_4; \sum_i p_i = 1)$. The pmf of $\boldsymbol{Y}$ is given by

$$p_{\boldsymbol{Y}}(\boldsymbol{y}) = \begin{cases} \frac{280!}{y_1!y_2!y_3!y_4!} p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4}, & y_i = 0, 1, ..., 280; \sum_i y_i = 280 \\ 0, & \text{otherwise.} \end{cases}$$

FACTS: If $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_k) \sim \text{mult}(n, p_1, p_2, ..., p_k; \sum_i p_i = 1)$, then

- The marginal distribution of Y_i is $b(n, p_i)$, for $i = 1, 2, ..., k$.
- $E(Y_i) = np_i$, for $i = 1, 2, ..., k$.
- $V(Y_i) = np_i(1 - p_i)$, for $i = 1, 2, ..., k$.
- The joint distribution of (Y_i, Y_j) is trinomial$(n, p_i, p_j, 1 - p_i - p_j)$.
- $\text{Cov}(Y_i, Y_j) = -np_ip_j$, for $i \neq j$.

4.11 The bivariate normal distribution

TERMINOLOGY: The random vector (Y_1, Y_2) has a **bivariate normal distribution** if its joint pdf is given by

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-Q/2}, & (y_1, y_2) \in \mathcal{R}^2 \\ 0, & \text{otherwise,} \end{cases}$$

where

$$Q = \frac{1}{1-\rho^2}\left[\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{y_1-\mu_1}{\sigma_1}\right)\left(\frac{y_2-\mu_2}{\sigma_2}\right) + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2\right].$$

We write $(Y_1, Y_2) \sim \mathcal{N}_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. There are 5 parameters associated with this bivariate distribution: the marginal means (μ_1 and μ_2), the marginal variances (σ_1^2 and σ_2^2), and the correlation $\rho \equiv \rho_{Y_1,Y_2}$.

FACTS ABOUT THE BIVARIATE NORMAL DISTRIBUTION:

1. Marginally, $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.
2. Y_1 and Y_2 are **independent** $\iff \rho = 0$. This is only true for the bivariate normal distribution (remember, this does **not** hold in general).
3. The conditional distribution
$$Y_1|\{Y_2 = y_2\} \sim \mathcal{N}\left[\mu_1 + \rho\left(\frac{\sigma_1}{\sigma_2}\right)(y_2-\mu_2), \sigma_1^2(1-\rho^2)\right].$$
4. The conditional distribution
$$Y_2|\{Y_1 = y_1\} \sim \mathcal{N}\left[\mu_2 + \rho\left(\frac{\sigma_2}{\sigma_1}\right)(y_1-\mu_1), \sigma_2^2(1-\rho^2)\right].$$

Exercise: Suppose that $(Y_1, Y_2) \sim \mathcal{N}_2(0, 0, 1, 1, 0.5)$. What is $P(Y_2 > 0.85|Y_1 = 0.2)$?
Answer: From the last result, note that, **conditional** on $Y_1 = y_1 = 0.2$, $Y_2 \sim \mathcal{N}(0.1, 0.75)$. Thus, $P(Y_2 > 0.85|Y_1 = 0.2) = P(Z > 1) = 0.1587$. Interpret this value as an area.

4.12 Conditional expectation

4.12.1 Conditional means and curves of regression

TERMINOLOGY: Suppose that X and Y are **continuous** random variables and that $g(X)$ and $h(Y)$ are functions of X and Y, respectively, Recall that the conditional distributions are denoted by $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$. Then,

$$\begin{aligned} E[g(X)|Y=y] &= \int_{\mathcal{R}} g(x) f_{X|Y}(x|y)dx \\ E[h(Y)|X=x] &= \int_{\mathcal{R}} h(y) f_{Y|X}(y|x)dy. \end{aligned}$$

If X and Y are **discrete**, then sums replace integrals.

IMPORTANT: It is important to see that, in general,

- $E[g(X)|Y=y]$ is a function of y, and
- $E[h(Y)|X=x]$ is a function of x.

CONDITIONAL MEANS: In the definition above, if $g(X)=X$ and $h(Y)=Y$, we get (in the continuous case),

$$\begin{aligned} E(X|Y=y) &= \int_{\mathcal{R}} x f_{X|Y}(x|y)dx \\ E(Y|X=x) &= \int_{\mathcal{R}} y f_{Y|X}(y|x)dy. \end{aligned}$$

$E(X|Y=y)$ is called the **conditional mean** of X, given $Y=y$; it is the mean of the conditional distribution $f_{X|Y}(x|y)$. On the other hand, $E(Y|X=x)$ is the **conditional mean** of Y, given $X=x$; it is the mean of the conditional distribution $f_{Y|X}(y|x)$.

Example 4.23. In a simple genetics model, the proportion, say X, of a population with Trait 1 is always less than the proportion, say Y, of a population with trait 2. In Example 4.3, we saw that the random vector (X,Y) has joint pdf

$$f_{X,Y}(x,y) = \begin{cases} 6x, & 0<x<y<1 \\ 0, & \text{otherwise.} \end{cases}$$

In Example 4.5, we derived the conditional distributions

$$f_{X|Y}(x|y) = \begin{cases} 2x/y^2, & 0 < x < y \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad f_{Y|X}(y|x) = \begin{cases} \frac{1}{1-x}, & x < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the conditional mean of X, given $Y = y$ is

$$\begin{aligned} E(X|Y=y) &= \int_0^y x f_{X|Y}(x|y)dx \\ &= \int_0^y x\left(\frac{2x}{y^2}\right)dx = \frac{2}{y^2}\left(\frac{x^3}{3}\bigg|_0^y\right) = \frac{2y}{3}. \end{aligned}$$

Similarly, the conditional mean of Y, given $X = x$ is

$$\begin{aligned} E(Y|X=x) &= \int_x^1 y f_{Y|X}(y|x)dy \\ &= \int_x^1 y\left(\frac{1}{1-x}\right)dy = \frac{1}{1-x}\left(\frac{y^2}{2}\bigg|_x^1\right) = \frac{1}{2}(x+1). \end{aligned}$$

That $E(Y|X=x) = \frac{1}{2}(x+1)$ is not surprising because $Y|\{X=x\} \sim \mathcal{U}(x,1)$. □

TERMINOLOGY: Suppose that (X, Y) is a bivariate random vector.

- The graph of $E(X|Y = y)$ versus y is called the **curve of regression** of X on Y.

- The graph of $E(Y|X = x)$ versus x is called the **curve of regression** of Y on X.

The curve of regression of Y on X, from Example 4.23, is depicted in Figure 4.19.

4.12.2 Iterated means and variances

REMARK: In general, $E(X|Y = y)$ is a function of y, and y is fixed (not random). Thus, $E(X|Y = y)$ is a fixed number. However, $E(X|Y)$ is a function of Y; thus, $E(X|Y)$ is a random variable! Furthermore, as with any random variable, it has a mean and variance associated with it!!

ITERATED LAWS: Suppose that X and Y are random variables. Then the **laws of iterated expectation** and **variance**, respectively, are given by

$$E(X) = E[E(X|Y)]$$

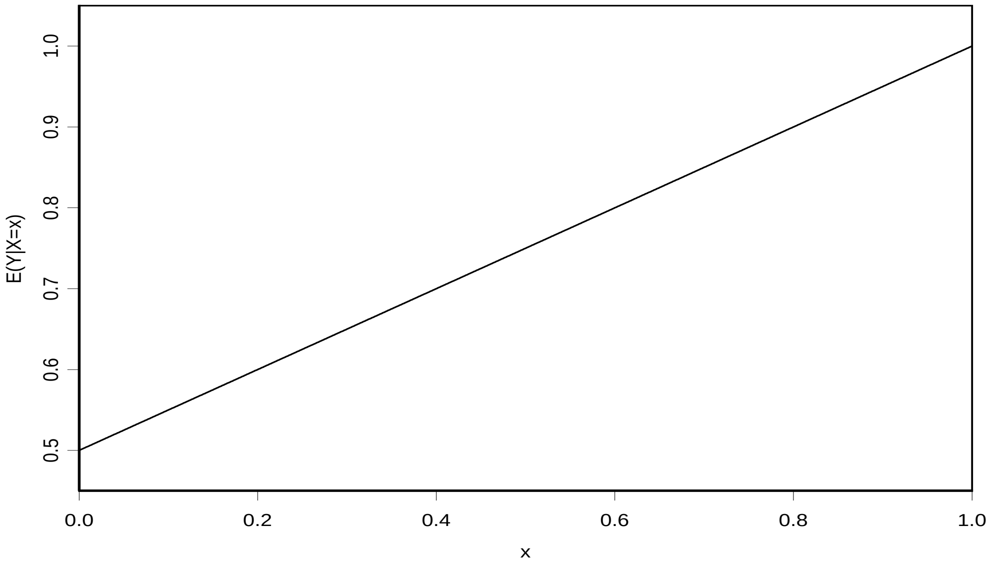

Figure 4.19: *The curve of regression* $E(Y|X=x)$ *versus* x *in Example* 4.23.

and

$$V(X) = E[V(X|Y)] + V[E(X|Y)].$$

NOTE: When considering the quantity $E[E(X|Y)]$, the inner expectation is taken with respect to the conditional distribution $f_{X|Y}(x|y)$. However, since $E(X|Y)$ is a function of Y, the outer expectation is taken with respect to the marginal distribution $f_Y(y)$.

Proof. We will prove that $E(X) = E[E(X|Y)]$ for the continuous case. Note that

$$\begin{aligned} E(X) &= \int_{\mathcal{R}}\int_{\mathcal{R}} x f_{X,Y}(x,y)dxdy \\ &= \int_{\mathcal{R}}\int_{\mathcal{R}} x f_{X|Y}(x|y)f_Y(y)dxdy \\ &= \int_{\mathcal{R}} \underbrace{\left[\int_{\mathcal{R}} x f_{X|Y}(x|y)dx\right]}_{E(X|Y=y)} f_Y(y)dy = E[E(X|Y)]. \ \square \end{aligned}$$

Example 4.24. Suppose that in a field experiment, we observe Y, the number of plots, out of n, that respond to a treatment. However, we don't know the value of p, the probability of response, and furthermore, we think that it may be a function of location, temperature, precipitation, etc. In this situation, it might be appropriate to regard p as a **random variable**! Specifically, suppose that the random variable P varies according to a beta(α, β) distribution. That is, we assume a **hierarchical structure**:

$$\begin{aligned} Y|P = p &\sim \text{binomial}(n, p) \\ P &\sim \text{beta}(\alpha, \beta). \end{aligned}$$

The (unconditional) mean of Y can be computed using the iterated expectation rule:

$$E(Y) = E[E(Y|P)] = E[nP] = nE(P) = n\left(\frac{\alpha}{\alpha+\beta}\right).$$

The (unconditional) variance of Y is given by

$$\begin{aligned} V(Y) &= E[V(Y|P)] + V[E(Y|P)] \\ &= E[nP(1-P)] + V[nP] \\ &= nE(P-P^2) + n^2V(P) \\ &= nE(P) - n\{V(P) + [E(P)]^2\} + n^2V(P) \\ &= n\left(\frac{\alpha}{\alpha+\beta}\right) - n\left[\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \left(\frac{\alpha}{\alpha+\beta}\right)^2\right] + \frac{n^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= n\left(\frac{\alpha}{\alpha+\beta}\right)\left[1 - \left(\frac{\alpha}{\alpha+\beta}\right)\right] + \underbrace{\frac{n(n-1)\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}_{\text{extra variation}}. \end{aligned}$$

Unconditionally, the random variable Y follows a **beta-binomial** distribution. This is a popular probability model for situations wherein one observes binomial type responses but where the variance is suspected to be larger than the usual binomial variance. □

BETA-BINOMIAL PMF: The probability mass function for a **beta-binomial** random variable Y is given by

$$\begin{aligned} p_Y(y) = \int_0^1 f_{Y,P}(y,p)dp &= \int_0^1 f_{Y|P}(y|p)f_P(p)dp \\ &= \int_0^1 \binom{n}{y} p^y(1-p)^{n-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}dp \\ &= \binom{n}{y} \frac{\Gamma(\alpha+\beta)\Gamma(y+\alpha)\Gamma(n+\beta-y)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)}, \end{aligned}$$

for $y = 0, 1, ..., n$, and $p_Y(y) = 0$, otherwise.